(Sub)-millimeter wave on-wafer calibration and device characterization

Marco Cabbia

université de BORDEAUX

THÈSE PRÉSENTÉE
POUR OBTENIR LE GRADE DE

DOCTEUR DE L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE SCIENCES PHYSIQUES ET DE L'INGÉNIEUR
Specialité Électronique

Par Marco CABBIA

(Sub-)Millimeter Wave On-Wafer Calibration and Device Characterization

Sous la direction de : Thomas ZIMMER
Co-encadrée par : Sébastien FREGONESE
Co-encadrée par : Marina DENG

Soutenue le 15/01/2021

Membres du jury :

M. FERRARI Philippe	Professeur	Université de Grenoble-Alpes	Président
M. GAQUIÈRE Christophe	Professeur	Université de Lille	Rapporteur
M. SPIRITO Marco	Professeur associé	TU Delft	Rapporteur
M. DURAND Cédric	Ingénieur	STMicroelectronics	Invité
M. YADAV Chandan	Professeur assistant	NIT Calicut	Invité
M. FREGONESE Sébastien	Chargé de recherche	CNRS	Co-encadrant
Mme DENG Marina	Maître de conférences	Université de Bordeaux	Co-encadrante

Calibrage sur plaquette et caractérisation des dispositifs à ondes (sub-)millimétriques

Résumé : Les mesures de précision jouent un rôle crucial dans l'électronique, en particulier dans la caractérisation des transistors bipolaires à hétérojonction (HBT) à base de silicium embarqués dans des dispositifs pour applications THz utilisant la technologie BiCMOS. Grâce aux innovations récentes en ce qui concerne la fabrication de technologies à l'échelle nanométrique, les dispositifs capables de fonctionner dans la région des ondes submillimétriques deviennent une réalité et doivent répondre à la demande de circuits et de systèmes haute-fréquence. Pour disposer de modèles précis à de telles fréquences, il n'est plus possible de limiter l'extraction des paramètres en dessous de 110 GHz, et de nouvelles techniques permettant d'obtenir des mesures fiables de dispositifs passifs et actifs doivent être étudiées. Dans cette thèse, nous examinerons la caractérisation des paramètres S sur silicium (on-wafer) de différentes structures de test passives et des HBT SiGe en technologie B55 de STMicroelectronics, jusqu'à 500 GHz. Nous commencerons par une introduction de l'équipement de mesure habituellement utilisé pour ce type d'analyse, puis nous passerons aux différents bancs de mesure adoptés au laboratoire IMS, et enfin nous nous concentrerons sur les techniques de calibrage et d'épluchage (de-embedding), en passant en revue les principales criticités de la caractérisation haute-fréquence et en comparant deux algorithmes de calibrage on-wafer (SOLT et TRL) jusqu'à la bande WR-2.2. Deux cycles de production de photomasques pour la caractérisation on-wafer, tous deux conçus à l'IMS, seront présentés : nous introduirons un nouveau design du floorplan et évaluerons sa capacité à limiter les effets parasites ainsi que l'effet de son environnement (substrat, structures voisines et diaphonie). Pour notre analyse, nous nous appuierons sur des simulations électromagnétiques et des simulations EM mixtes de modèle compacte + sonde, toutes deux incluant les modèles des sondes pour une évaluation des résultats de mesure plus proche des conditions réelles. Enfin, nous présenterons quelques structures de test pour évaluer les impacts indésirables sur les mesures d'ondes millimétriques et de nouvelles solutions de conception de lignes de transmission. Deux designs prometteurs seront soigneusement étudiés : le "layout M3", qui vise à caractériser le DUT dans un étalonnage à un seul niveau, et les "lignes à méandre", qui maintiennent la distance entre les deux sondes constante en évitant tout déplacement pendant les mesures sur silicium.
Mots-clés : Caractérisation, Lignes de transmission, Térahertz, Ondes millimétriques, Calibrage sur silicium, TBH en SiGe

---

(Sub-)Millimeter Wave On-Wafer Calibration and Device Characterization

Abstract: Precision measurements play a crucial role in electronic engineering, particularly in the characterization of silicon-based heterojunction bipolar transistors (HBTs) embedded into devices for THz applications using the BiCMOS technology. Thanks to ongoing innovations in terms of nanoscale technology manufacturing, devices capable of operating in the sub-millimeter wave region are becoming a reality, and need to support the demand for high frequency circuits and systems. To have accurate models at such frequencies, it is no longer possible to limit the parameter extraction below 110 GHz, and new techniques for obtaining reliable measurements of passive and active devices must be investigated. In this thesis, we examine the on-wafer S-parameters characterization of various passive test structures and SiGe HBTs in STMicroelectronics' B55 technology, up to 500 GHz. We start with an introduction of the measuring equipment usually employed for this type of analysis, then moving on to the various probe stations adopted at the IMS Laboratory, and finally focusing on calibration and deembedding techniques, reviewing the major criticalities of high-frequency characterization and comparing two on-wafer calibration algorithms (SOLT and TRL) up to the WR-2.2 band. Two photomask production runs for on-wafer characterization, both designed at IMS, are considered: we introduce a new floorplan design and evaluate its ability to limit parasitic effects as well as the effect of the environment (substrate, neighbors, and crosstalk). For our analysis, we rely on electromagnetic simulations and joint device model + probe EM simulations, both including probe models for an evaluation of measurement results closer to real-world conditions. Finally, we present some test structures to evaluate unwanted impacts on millimeter wave measurements and novel transmission line design solutions. Two promising designs are carefully studied: the "M3 layout", which aims to characterize the DUT in a single-tier calibration, and the "meander lines", which keeps the inter-probe distance constant by avoiding any sort of probe displacement during on-wafer measurements.
Keywords: Characterization, Transmission Lines, THz, Millimeter-Wave, On-Wafer Calibration, SiGe HBT

UMR 5218 - IMS - Laboratoire de l'Intégration du Matériau au Système
351 Cours de la libération, 33405 Talence CEDEX, France.

Thanks

A COMPLETE presentation of this three-year work on transistor characterization with extensive on-wafer calibration results can be found in the following of the manuscript. The author, however, would like to address here a few very personal words (actually, quite a lot) to thank all the people who meant for him the most during this Bordeaux journey, as well as to those who have been by his side long before it.

Tout d'abord, je tiens à remercier mon directeur de thèse, Monsieur Thomas Zimmer, dont l'expérience et la connaissance du sujet ont fait la personne de référence de mon travail. Tu as réussi à établir un équilibre sain entre rigueur et bienveillance ; je retiendrai surtout tes compétences de médiateur et ta capacité à relativiser les difficultés en gardant la tête froide et le moral de tout le monde. L'opportunité de collaborer avec l'équipe Modèle de l'IMS est venue presque par hasard, lors de mon expérience Erasmus que j'avais tant désirée. Pour m'avoir donné cette opportunité, je tiens à remercier Madame Cristell Maneux, qui a pris contact avec moi avant mon arrivée à Bordeaux et qui a encadré mes travaux de stage, et Monsieur Sébastien Fregonese, qui m'a guidé pendant le stage mais qui est devenu, surtout, l'un de mes principaux collaborateurs pendant cette thèse. Merci de m'avoir aidé à donner une direction précise à mes recherches, de m'avoir soutenu dans l'apprentissage et m'avoir offert toujours ton point de vue original, basé sur la rigueur scientifique et la franchise. Enfin, un immense merci à Madame Marina Deng, qui a cru en mes capacités et mes résultats dès le départ ; pour m'avoir poussé à y croire aussi. Tu as guidé ma thèse avec de solides compétences pédagogiques, offrant toujours des conseils scientifiques valables et une écoute sincère. Merci pour les nombreuses conversations que nous avons eues, qui m'ont permis de découvrir une forte proximité d'idées et de valeurs. Enfin, le voyage pour la conférence à San Antonio (et les boissons locales !) sont l'un des souvenirs les plus agréables et ineffaçables de ce doctorat.

I wish to express my gratitude to Chandan Yadav, who was my mentor in the early days of the PhD, and guided me day by day in my first steps into research with assertiveness, pushing me to constantly improve. I personally recall your genuine humbleness, altruism and kindness, proved to anyone, anytime, without any sort of affectation. A Madame Magali De Matos, pour l'infinie patience dont elle a fait preuve à chaque répétition des contacts de sondes, lors des campagnes de mesures, et pour m'avoir aidé à interpréter les (nombreux) résultats.

Besides my PhD advisors and collaborators, I would like to thank all the members of my thesis committee, Profs Philippe Ferrari, Christophe Gaquière and Marco Spirito, and Mr. Cédric Durand for kindly devoting their time to evaluating my work. I am deeply honored that you accepted to attend my PhD defense, remotely and even in-person, during these hard pandemic times we go through.

I am also very grateful to Karthi for the stimulating scientific talks we have had in these recent months. To Soumya, Abhishek and Bishwadeep for the fruitful collaborations and the cheerful atmosphere they have created in the student's room. Je tiens à remercier tous les autres doctorants, Marine, Ming Ming, Djeber, Mathieu J., Olivia, Florent, Quentin, Isabel, Jean-Baptiste, Ghyslain, Adrien, Mathieu M., pour l'ambiance conviviale et décontractée qu'ils ont instaurée. Je n'ai pas

toujours été capable de m'exprimer ouvertement mais je vous apprécie tous beaucoup, je garde de très bons souvenirs des moments que nous avons passés ensemble, spécialement nos aventures en montagne. Enfin, une pensée affectueuse aux amis de la BEE Branch et aux concurrents de « Ma thèse en 180 secondes ».

Je voudrais maintenant remercier ceux qui, hors du laboratoire et du travail de recherche, m'ont fait vivre pleinement ces années. A Camille et Antoine, mes amis voisins de l'étage, les premières personnes que j'ai rencontrées le soir de mon arrivée en France, il y a quatre ans. Vous m'avez accueilli comme quelqu'un de la famille et votre hospitalité et bonne humeur n'ont jamais faibli depuis. Aux voisins « au sens large », Elisa, Giacomo et Hugo, qui font partie des amis qui ont rendu mes premiers mois à Bordeaux si mémorables et qui restent proches de moi, malgré le temps et la distance.

Je tiens à remercier Saphia, pour ta positivité et ta curiosité. Pour toutes nos conversations amusantes et passionnantes, pour tes recommandations musicales et nos séries... même lorsque elles déçoivent après la saison un. A Marion, avec qui l'amitié a tant évolué depuis ta réponse à mon annonce sur la Carte des Colocs. Si je ne t'avais pas rencontré, beaucoup de choses aurait été sans doutes différentes. Il m'aurait manqué la personne sur laquelle compter, me donnant toujours des conseils utiles et des opinions valables. Celle qui, bien qu'étant seul dans un pays étranger, ne m'a pas abandonné mais qui, au contraire, m'a intégré à sa vie. D'ailleurs, si je peux exprimer tout cela en bon français, c'est notamment grâce à toi. J'ai hâte de trinquer à nouveau ensemble autour d'un Fernet de la Cueva et de reprendre nos voyages. Por último, un gracias muy especial a Lorena. Es raro conocer a alguien con quien tener una conexión tan inmediata y un diálogo tan espontáneo, basado en la confianza y la sinceridad, como entre nosotros; siempre presente para una reflexión franca sobre cómo nos sentimos... ¡Comiendo juntos un bol de yogur con cereales! Tu extroversión y tu fidelidad a ti misma constituyen la energía que transmites a las personas y fueron lo más significativo que aprendí en estos años. Sin mencionar el haber descubierto que quiero un gato.

Desidero poi ringraziare Giorgia. Ai nostri giri in bici a Dolo, in provincia di Bordeaux, ma anche, certamente, alle infinite pause pranzo nel campus e alle innumerevoli conversazioni... mature e intelligenti che abbiamo avuto insieme: degne di due dottori. Grazie a Lorenzo per avermi fatto scassare così tanto. Non vedo l'ora di tornare a spingere in bici e sulla tavola da surf assieme. Sono grato a entrambi perché ho passato con voi i mesi più spensierati, sentendo di potermi esprimere in tutta libertà. Spero mi abbiate perdonato per il ritardo quella mattina; alla fine, il weekend a San Sebastian non è stato così male, anzi, resta uno dei più bei ricordi che ancora oggi ho.

Quiero decirle gracias a tod@s mis amig@s barbud@s, sobre todo a Anto, Héctor, Rafa, Manu, Hugo y Arturo, por poder contar con ellos en tener un plan para una fiesta tranquila en la coloc... que siempre termina degenerando. Gracias por haberme alegrado los fines de semana en estos años (¡y que sean muchos años más!), por vuestras sorpresas y vuestro apoyo en estos meses escribiendo la tesis.

Anche se da anni sparpagliati per l'Europa, ridotti a vederci sotto le feste o a incastrare una mezza giornata attorno a Ferragosto, ci tengo a ringraziare tutti coloro per cui il piano tariffario è ben speso. Ad Andrea, Luca, Irene e Francesco, per esserci ancora praticamente ogni giorno, nonostante siamo lontani dal DEI, dal Paolotti, dalla Piovego e dal treno delle 7.18 (o 8.54, morbido). Grazie di motivarmi e ascoltare e rielaborare con pazienza anche le mie paranoie più astratte. Al mio compare, confidente e, senza dubbio, futuro socio in affari, Nicolò. Grazie per l'energia che mi trasmetti, per essere schietto e sempre presente per una chiamata e una risata, da anni; il valore che dai tu all'amicizia è l'esempio migliore che io abbia mai ricevuto da qualcuno. Ringrazio poi Lorenzo e Tommaso. È bello sapere che ormai siete a vostro agio qui da me, così tanto da voler persino dormire su un materasso in giardino. L'epidemia passerà presto

e, vedrete, saremo di nuovo 4 su 4. Grazie a Lisa per il suo appuntamento (quasi...) giornaliero con l'informazione la mattina. Mi ritengo fortunato perché tornare in Italia e riprendere le nostre conversazioni come se non fossi mai partito è una cosa veramente pazzesca. Ringrazio quindi Ilaria; nonostante tu risponda ancora ai miei messaggi per monosillabi e in ritardo, non hai mai avuto nessun problema a dirmi ciò che pensi e trovo in te, ancor oggi a distanza di 15 anni, un'interlocutrice sincera e tanto affine a me. E comunque: anche oggi si scopre domani... ma almeno oggi si è scritta una tesi.

Un grazie infine alla mia famiglia. Ai miei genitori, che malgrado le circostanze difficili degli ultimi anni hanno favorito in tutti i modi, da sempre e con convinzione, la mia formazione e la mia educazione. Grazie per appoggiarmi in ciò che mi gratifica di più: il mio percorso in Francia, nonostante la durezza della separazione. Ringrazio mia sorella, per l'ammirazione che mi dimostra; sono convinto che la tua personalità e le tue doti continueranno a portarti lontano. Grazie nonna, perché devo molto a te e alla zia, e non vi dimentico un solo giorno. Grazie zia Sabrina e zio Sandro per il vostro appoggio, i momenti divertenti e i viaggi che riprenderemo presto. Grazie Luca, Francesca, zio Antonio e zia Stefania, per seguirmi e incoraggiarmi con entusiasmo e affetto.

Mon doctorat se termine en cette année 2020, qui a été véritablement une drôle d'année. J'ai été contraint de m'isoler pendant de longues périodes pour achever l'écriture de ce manuscrit, certes, mais aussi pour échapper au virus, comme pour la plupart d'entre nous. Cela m'a inévitablement éloigné de tous ceux que j'ai remercié ci-dessus. Cependant, une personne est restée près de moi, suivant mes horaires irréguliers, durant les confinements et lorsque je rédigeais. Cette thèse est en grande partie le résultat de ses encouragements incessants, de son soutien concret dans les tâches quotidiennes, de son écoute et ses gestes dévoués envers moi. A tes vertus et tes qualités que tu sous-estimes parfois mais que j'admire ; la complexité de tes émotions, ton âme généreuse et joyeuse. Pour la ferme confiance en notre couple. Enfin, simplement, pour l'amour et le bonheur. Merci Déborah. Prends tes affaires... car nous partons en voyage bientôt !

Al me Burici.
L'è duda ancia chista, a'to vist?

"Someday soon, you're gonna have families of your own and if you're lucky, you'll remember the little moments like this, that were good."

Contents

Chapter 1

Introduction

IN THIS DISSERTATION we will discuss on-wafer calibration approaches and general device characterization techniques for millimeter and sub-millimeter wave frequency transistors. The field of THz integrated circuit technologies has grown tremendously over the past decade. Recent research has focused on devices with different substrates (III-V and silicon-based semiconductors) and different transistors (field-effect and heterojunction bipolar devices) to bring multiple new functions in the fields of communication, imaging and sensing. We will focus on silicon-germanium heterojunction bipolar transistors (SiGe HBTs), and we will try to explore new approaches to characterize them.

We will begin in Chapter One with a brief introduction on the advantages of these innovative applications and the physical mechanisms that govern HBTs and the most usual high frequency performance metrics for these devices. We will briefly present the manufacturing process of the bipolar CMOS (BiCMOS) technology that is adopted for the entire thesis, putting into the context of the state of the art.

In Chapter Two, we will introduce the useful concepts for understanding the measurement process (S-parameters) carried out by a vector network analyzer (VNA), normally used on measurements of electromagnetic signals. We will describe its main architectural features, and thus we will be able to introduce the problems posed by measurement uncertainties. We will define error models and techniques (algorithms) usually employed to remove error terms. Of these, two will later be resumed and performed: SOLT and TRL calibrations. We will then talk about calibration on silicon supports (on-wafer), a substrate chosen to minimize the discontinuities of the measurement environment and the challenges that arise in the characterization at millimeter and sub-millimeter frequencies. Our measurement setup, consisting of VNA, connections and probes will finally be introduced.

In Chapter Three, we will move on to the actual measured devices, presenting the wafer (with two different layout approaches) where our test structures –the HBTs and the calibration standards– lie. We will describe the characteristics of the back-end-of-line (BEOL) and the properties of each calibration and verification standard, highlighting the important changes of our second production run that allow a better calibration (and therefore of the measurements) quality. The layout of our devices and transmission lines will be put in context with current trends by other laboratories and research centers. We will also detail our simulation setup for verifying the measurements taken. We will talk about our calibration "toolkit" and evaluate the effectiveness of the impedance correction that must always be used after TRL calibration. Our measurements calibrated with SOLT and TRL will be compared up to 500 GHz. Finally we will compare the two production runs and the different properties, also evaluating, for the most recent run, the ability to reduce the effect of adjacent structures.

Chapter Four will be devoted to presenting several innovative approaches to on-wafer calibration structures. Specifically, two will be treated in detail. The first, named "M3 layout" claims to avoid the classic two-tier calibration, which uses two successive steps to complete the removal of the contribution from the measurement environment and the BEOL. The microstrip lines built

at the metal 3 level of the BEOL allow to avoid the second calibration step, given the proximity to metal 1 level, where the transistor to be measured is located. The ability to obtain quality measurements with this technique will be evaluated, comparing with the standard approach and some variants. A calibration called "3D TRL", which brings the single calibration step directly to the transistor level, will be briefly shown, as a promising alternative to the M3 layout. The second approach concerns the "meander layout": we have designed some microstrip lines that are not straight but present snake-shaped signal track. With this design we hope to avoid the natural horizontal probe distancing when measuring long transmission lines, which are required to perform TRL calibration. Given the non-univocal definition of the length of these lines, techniques for defining an effective length of the meander lines are necessary and will be presented. The results and limitations of this design will once again be confronted with a classic approach. Finally, we will take a look at the third production run, already designed but yet to be measured.

Chapter Five will sum up and draw conclusions on the different on-wafer characterization techniques of HBTs up to 500 GHz.

The research goal that will be directly addressed in this manuscript is to provide on one hand a complete benchmarking of fully on-wafer measurements and calibration techniques for both passive test structures and transistor from DC to 500 GHz, with EM simulations backing our conclusions on every part of the spectrum, and on the other to propose possible calibration standard implementations expressly design for millimeter-wave measurements.

1.1 Millimeter Wave and Terahertz Radiation

Fig. 1.1 shows a representation of the electromagnetic spectrum where the bands are shown as a function of frequency, wavelength and energy. At the lower end of the picture is the radio spectrum, a frequency region – from few kHz to few GHz – where classical radio systems usually work (FM, AM, cellular radio, etc...). The other side is the domain of optical radiation bands (around 10^{13} Hz) and it extends up to gamma rays (around 10^{21} Hz). Optoelectronics deals with light considered as both visible and invisible radiation (photodiodes, lasers, optical fibers, etc...).

Millimeter waves (also abbreviated as *mmW* or *MMW*) and terahertz radiation (also called sub-millimeter radiation and abbreviated as *T-ray* or simply *THz*) cover the region of the spectrum from microwave to optical frequencies, the so called "terahertz gap". More specifically, the range for the millimeter band is 30-300 GHz (equivalent to λ = 1-10 mm in vacuum), while the range for the terahertz band proper is 300 GHz to 3 THz (equivalent to λ = 100 µm to 1 mm in vacuum) [86]. This regime carries a lot of the benefits of both sides: like radio waves it can penetrate through a variety of non-conducting objects and walls, and like visible light it has very short wavelengths which give very precise measurements and high quality images (Fig. 1.2). Terahertz radiation has limited penetration through fog and cloud and cannot penetrate liquid water and metal [75].

At the radio side of the spectrum (microwaves) we typically use electronic devices and the power available decreases at higher frequency; while for the upper side of the spectrum (infrared light waves), photonic devices are used, in which the energy per photon decreases at lower frequencies, and so does the available power for these devices. In the THz region, the frequency of electromagnetic radiation becomes too high to be measured by directly counting cycles through an electronic counter, and, similarly, in this range the generation and modulation of coherent electromagnetic signals is no longer possible by conventional radio-frequency and microwave electronic devices [75].

THz systems research has taken a major leap forward from laser-based technologies for generation and detection of sub-millimeter signals, where femtosecond lasers combined with ultrafast lightwave-to-THz converter are employed. However, this optical systems are bulky and expensive and the new systems could benefit from integrated terahertz circuits. Nowadays, the aim is no more to close the "terahertz gap", but to do it in meaningful ways. Much of the recent effort has been dedicated to conceive technologies compatible or realized in solid-state semiconductor technology, operating at room temperature and at low cost. The high level of integration of integrated circuits will indirectly increase the achievable complexity and reconfigurability of the systems, including electronic reconfigurability of the wavefront and polarization of the THz fields [108].

There are some unique specifications that make terahertz waves attractive for the scientific community. Primarily for spectroscopy, allowing the investigation of composition and physical structure of matter in the fields of chemistry, physics and astronomy. As a matter of fact, for stimulating a transition between energy states in order to measure the spectrum of molecules (e.g. in the gas state), rotational and vibrational frequencies are considered: many of those frequencies lie in the millimeter and sub-millimeter region. Also, many optically opaque materials are

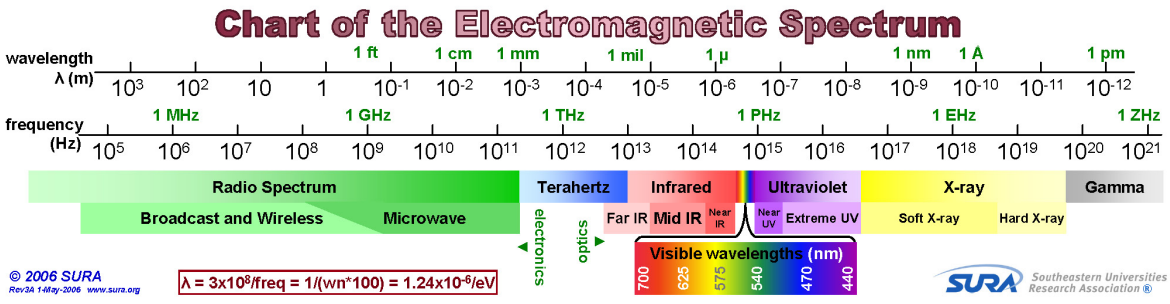


Figure 1.1: The electromagnetic spectrum (after [110]).

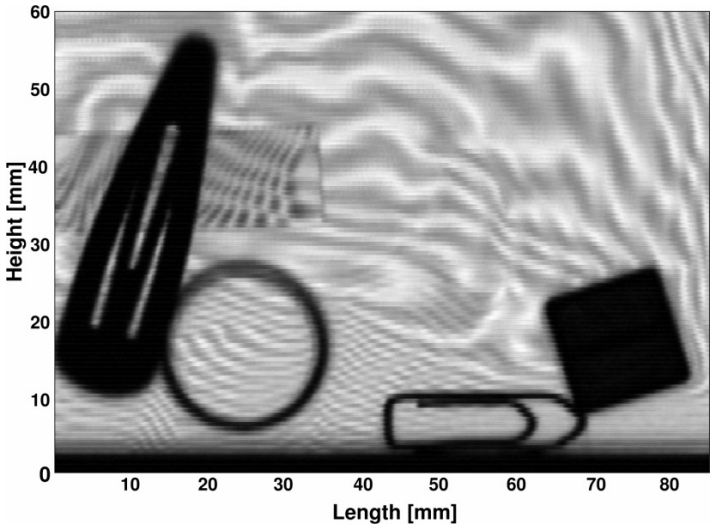


Figure 1.2: Captured 0.6-THz image of a postal envelope, revealing its hidden office supplies (such as paper clips, tape and candy bars). (after [76]).

more transparent to terahertz frequencies, and this allows to see through these objects and access environments that are usually inaccessible. Terahertz wave energy is much less than X-rays and even optical waves (photons at 1THz have an energy of just 4.1 meV [76]). As a result, they are non-ionizing, non-destructive, and they can be used for different medical imaging and sensing applications, as they have less chances of damaging tissues and causing destruction in products.

THz capabilities on imaging and sensing can be applied for non-destructive quality control, radar, robotics and automation and will be achievable in the near future and even nowadays for some application in the millimeter wave range. At THz frequencies, hybrid silicon-/III-V compounds -based solutions will probably be imagined.

1.1.1 Some Applications

The electronic devices that can reach such high frequencies and high powers are mainly used to fabricate monolithic microwave integrated circuits (MMICs) that perform the function of power amplifiers, low-noise amplifiers (LNAs), flash ADCs and voltage-controlled oscillators. It has already been mentioned that the panorama of possible applications is vast.

For example, in medical imaging, a skin cancer image obtained by a THz camera (Fig. 1.3) gives better contrast compared to classical optical imagining because of the strong absorption by water at these specific wavelengths [14]. Also, in many early stages of organic material decay when physical damage has not happened yet, classic imaging cannot see any kind of erosion even if actually a transformation is happening. Though since this decaying material has a different water content, THz waves prove to be reliable in detecting them.

Much like in the commonly experienced airport security screenings, which allow to find potentially dangerous carried objects by seeing through clothing and luggage, many screenings can be made by devices working at millimeter waves. By extending these technologies to higher frequencies, it can be possible to take higher resolution images – even from a large distance – by also applying higher powers.

In the pharmaceutical industry, there are two main issues tied up with medicine production. When a capsule is produced, for instance, one wants to make sure that the right thickness of coating is used because this will affect how long it takes to the drug to be released into the body. THz waves are good in both penetrating the capsule enough to measure the thickness and at the same time preserving its contents. The other problem is that the chemical inside the drug tends to crystallize in different forms, which again affects its solubility in water and its release time into the body, and with many classical chemical sensors one is not able to sense that. However, THz waves

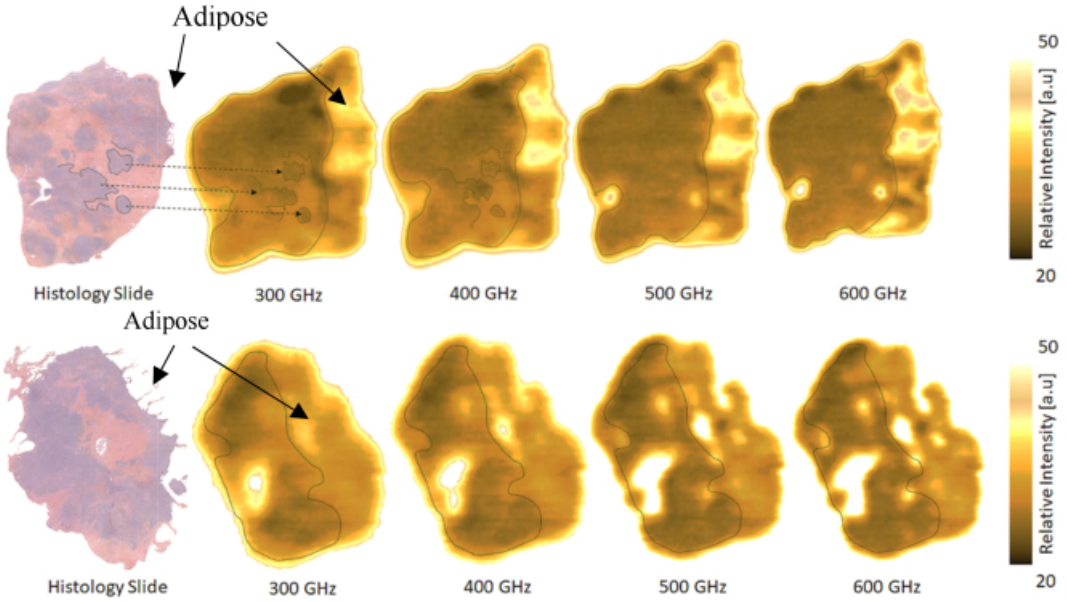


Figure 1.3: Comparison between histology slides and corresponding THz-images at 300, 400, 500 and 600 GHz. Frequencies ranging from 300 GHz to 400 GHz exhibit demarcations between cancerous regions and healthy fibers (from [15]).

can identify different rotational and vibrational frequencies of molecules and therefore spot in a non-invasive way if there are wrong crystalline forms.

For industrial quality control, there exist many applications to see if the right thickness of a material is used. In food industry, it's possible to determine if crops and grains quality is good and if they are fresh.

In safety systems for automotive, many radars are needed to accurately track the movement of objects all around the vehicle, opening the way to fully autonomous cars with cameras able to detect obstacles and moving objects under poor visibility and with a quick response.

Tremendous progress has been made in THz imagers, including 1,000 pixels THz video-rate integrated camera and active image systems. Sub-millimeter waves are limited in long-distance communications; however, at shorter distances (within 10 m) this band may still allow high bandwidth wireless networking systems, especially indoor systems, blazing a trail to next generations of standards for broadband cellular networks (5G+) and Internet of Things (IoT) networks.

Though, the very first type of application where THz waves were applied was space and atmospheric studies: THz sensors have been flying for years now within Earth-observing satellites to detect specific spectral lines of gases which determine the health of the Earth atmosphere and the ozone layer. Nevertheless, these satellites are carrying very large and heavy lasers; by imagining to develop even more these technologies in the future, one will get more compact and more performing sources to fit in a single satellite.

1.2 Bipolar and BiCMOS Transistors

Even if the usage of MOSFETs in the industry of semiconductors has been predominant for the last three decades, bipolar transistors have been vastly used in the past and thanks to improvements in bipolar technologies, they are currently maintaining and extending their predominance in many circuit applications, notably in high-speed performing systems.

The bipolar structure provides several natural advantages, such as: a short transit time for electrons flowing from emitter to collector (higher cut-off frequency); higher output current due to electron flowing through the entire emitter area (and not just a channel); direct control of the

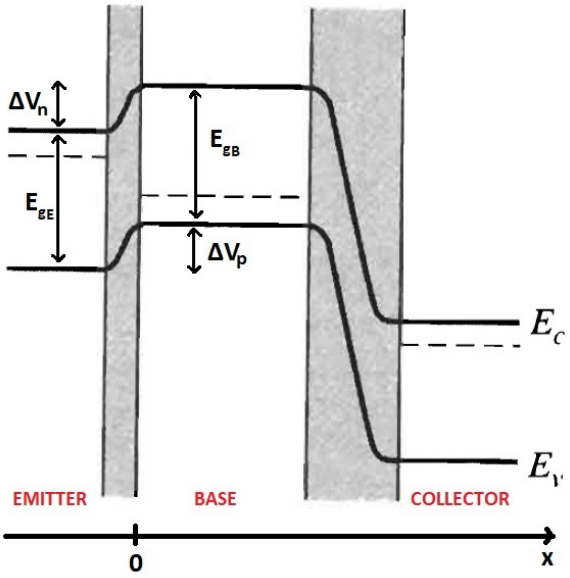


Figure 1.4: Energy-band diagram of a npn homojunction bipolar transistor.

output current through the input voltage leading to high transconductance; turn-on voltage independent of size; possibility of working either with high and low currents without experiencing considerable delays [111].

In Fig. 1.4 the energy-band diagram of a BJT along the direction of the electrons' travel is shown. Its configuration shows the forward active mode, that is the emitter-base junction (EBJ) is forward biased while the base-collector junction (CBJ) is reverse biased. In this way, electrons can surmount the EBJ barrier potential and they are swept through the CBJ by the strong electric field inside the space charge region (SCR).

It can be derived [72, 111] the forward DC current gain of bipolar transistors, which is defined as the ratio between the two currents, namely $\beta_0 = I_C/I_B$ as:

$$\beta_0 = \kappa \exp\left(\frac{\Delta E_g}{q V_T}\right) \tag{1.1}$$

where $\Delta E_g = E_{gE} - E_{gB} = q(\Delta V_p - \Delta V_n)$ is the emitter-base bandgap difference, V_T is the so-called *thermal voltage*, and κ is defined as:

$$\kappa = \frac{D_B}{D_E} \cdot \frac{n_{Eo}}{p_{Bo}} \cdot \frac{L_E}{W_B} \tag{1.2}$$

and takes into account all the contributions due to the diffusion of carriers, the concentration and the geometry of both base and emitter. Some general considerations can be done.

Firstly, the ratio between n_{Eo} and p_{Bo} is to keep high, in order to maintain κ high, or equivalently get high current gain. So that is why for a Si homojunction transistor, for example, emitter doping levels of $10^{20}\,cm^{-3}$ and base doping levels on the order of 10^{17}-$10^{18}\,cm^{-3}$ are typically used. Further, Eq. 1.1 motivates why the base width is kept thin: κ increases for small W_B.

Though the strongest dependence of β_0 stems from the argument of the exponential, ΔE_g. In a classical BJT (homojunction transistor) the effective difference between bandgaps is approximately zero, but it can even be a small negative value [111]. Because of the heavy doping of the emitter, it has been proved both theoretically and experimentally [58] that the bandgap on the emitter side even shrinks according to the emitter doping concentration and the gain is reduced. HBTs can provide a positive bandgap difference.

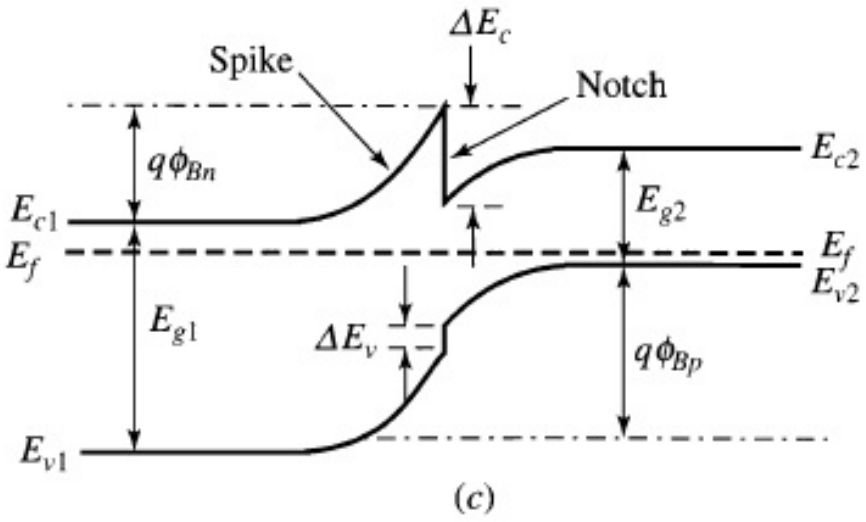


Figure 1.5: A spike appears in the emitter-base heterojunction when different materials are put into contact (after [123]).

1.2.1 Heterojunction Bipolar Transistors

The idea of bandgap engineering dates back to the 1950s and was first conceived by Shockley and Kroemer [56]. They noticed that a dramatic improvement of the current gain in bipolar devices could have been achieved by creating a much wider bandgap for the emitter compared to the base. However, the technology of *heterojunction bipolar transistors* (HBTs) was developed decades later and this delay was due to the struggle in dealing with contacting interfaces from different materials free of defects and imperfections, which usually come along with lattice mismatch.

Eq. 1.1 showed how the current gain is affected by a possible bandgap difference. In heterojunctions, the change in energy gap is made of two steps that arise at the conduction band and valence band respectively, that is $\Delta E_g = \Delta E_c + \Delta E_v$ (Fig. 1.5).

By working on the association of materials in HBTs, ΔE_g is typically chosen to be greater than 250 meV and β_0 is 10^4 times greater than the homojunction case for the same doping profile. A higher level of doping of the emitter compared to the base is no longer necessary; instead, p_{Bo} can be risen up to $10^{20}\,cm^{-3}$, thus reducing the base resistance. High gain and narrow base are still maintained. On a standard BJT, this move would have drastically reduced β_0: such a trade-off is removed. Similarly, n_{Eo} can be reduced, thus increasing the SCR on the emitter side and equivalently reducing the emitter junction capacitance [111].

Putting in contact regions with different bandgaps gives rise to discontinuities between the conduction bands and the valence bands as shown in Fig. 1.5. We can see that a spike may appear in front of the path of the electrons flowing from emitter to base (for example in some III-V materials), eventually reducing the injection efficiency of the carriers (abrupt junction). Grading the junction over several hundreds angstroms may solve this issue.

Fig. 1.6 shows what happens if we extend the concept of heterojunctions to the CBJ too and we increase the collector bandgap. When the CBJ is forward biased (saturation mode), fewer holes flow from the base into the collector and storage in the collector decreases, resulting in a quicker turn-off of the device. In the case of double heterojunction transistors, grading becomes mandatory in order not to suppress collector current [111].

1.2.2 High Frequency Performance

Two are the main figures of merit for designing high frequency circuits: *current-gain cutoff frequency* f_T and the *maximum frequency of oscillation* f_{max} [111, 90]. However, in systems where

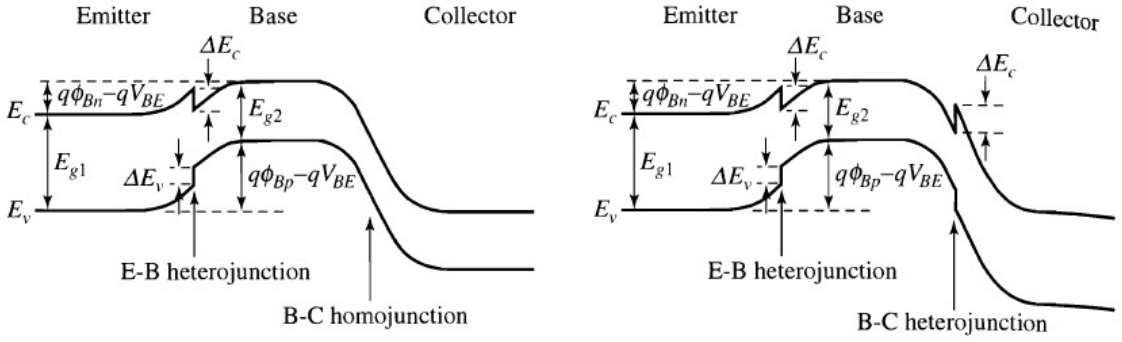


Figure 1.6: Single heterojunction transistor (at the left) and double heterojunction transistor (at the right) (after [123]).

high-frequency components – like HBTs – are used, f_T and f_{max} are far from providing an exhaustive characterization of the overall circuit. Amplifier noise figure, oscillator phase noise, 1/f noise corner, breakdown voltages (linked to the available power), thermal stability, etc... are to be taken into account when designing such circuits.

When considered solely by the transistor's metrics, InP-based HEMTs and HBTs display superior performances compared to CMOS, SiGe, and BiCMOS technology: InP-based HEMTs have reached 1.5 THz of f_{max} [69] while, due to limitations related to parasitics and contact resistances, silicon-based technology suffers from lower maximum frequency of oscillation (up to 0.72 THz [18]). For efficient THz signal generation and sensitive signal detection it is important to increase f_{max}, which also implies increasing the amplification range. Indeed, while III-V technologies exhibit power reaching the mW range beyond 1 THz, silicon-based phased arrays have demonstrated up to 0.1 mW at 1 THz, yet they provide a more valuable platform for massive integration [108].

Here below is an insight on these two important parameters, f_T and f_{max}.

Transit Frequency The incremental current gain β_F is not constant over frequency, we can approximate its behavior to a first-order system. In term of the *h*-parameters (see Appendix A), we can write:

$$\beta_F \triangleq \frac{di_C}{di_B} \equiv h_{21}(f) \triangleq \left.\frac{i_C}{i_B}\right|_{v_C=0} = \frac{\beta_0}{1+j2\pi f\,\tau_{EC}\beta_0} \tag{1.3}$$

where β_0 is the DC current gain (Eq. 1.1) and τ_{EC} is the emitter-to-collector transit time, closely linked to f_T. Indeed, this cut-off frequency is defined as the highest frequency at which the transistor current gain is equal to one, or more precisely, the frequency at which the incremental current gain equals one. So, at HF, we can rewrite:

$$|h_{21}(f)| = \frac{1}{2\pi f\,\tau_{EC}} = \frac{f_T}{f} \Longrightarrow |h_{21}(f)|\cdot f = f_T \tag{1.4}$$

where we have defined $f_T \triangleq \dfrac{1}{2\pi\tau_{EC}}$. We can also derive that:

$$\left|\frac{di_C}{di_B}\right| = \left|\frac{di_C}{dq_B}\cdot\frac{1}{j2\pi f}\right| = \frac{f_T}{f} \Longrightarrow \frac{1}{2\pi f_T} = \tau_{EC} = \frac{dq_B}{di_C} \tag{1.5}$$

where dq_B is the base charge associated with an increment of the input voltage and di_C is the collector current associated with an increment of the input voltage as well. We can therefore give a more physical definition of the emitter-to-collector transit time, relying on the fact that any variation of the bias point of the device is related to changes of the carrier densities within the device which are fed by currents into the device contacts:

$$\frac{1}{2\pi f_T} = \tau_{EC} = \frac{C_{diff}}{g_m} + \frac{C_{BE}}{g_m} + C_{BC}\left[\frac{1}{g_m} + (R_E + R_C)\right] \tag{1.6}$$

These parameters are related to those of the model presented in Fig. 1.9. Here, R_E and R_C are the emitter and collector resistances and g_m is the transconductance, which displays a dependence of f_T to the collector current. As for the capacitances, C_{BC} is the base-collector capacitance, composed of the intrinsic part, which is mainly a depletion capacitance in forward active mode, and of an extrinsic part, related to the base link region; the base-emitter capacitance C_{BE}, which includes the oxide capacitance of the junction as well as the diffusion capacitance, in forward active mode; the diffusion capacitance C_{diff}.

The diffusion capacitance accounts for the storage of locally compensated minority charges during forward active mode. This delay contribution is composed by:

$$\frac{C_{diff}}{g_m} = \tau_E + \tau_{EB} + \tau_B + \tau_{BC} \tag{1.7}$$

τ_E and τ_{EB} denote the delay due to storage in the emitter and base-emitter junction, respectively. While τ_{EB} can be significant in particular at high current densities, τ_E has less importance, since the amount of holes stored in the emitter is inversely proportional to the current gain. When designing HBTs, the emitter depth W_E should be as small as possible and the emitter doping as high as possible to maintain τ_E low.

τ_B account for the delay in the base and can be approximated by:

$$\tau_B = \frac{W_B^2}{2D_B} + \frac{W_B}{v_{sat}} \tag{1.8}$$

where W_B is the base region width and v_{sat} is the electron saturation velocity. Eq. 1.8 shows a second-order proportionality between τ_B and W_B, therefore explaining why it is essential for high-speed bipolar transistor to come with very small base widths.

τ_{BC} is the time in which electrons travel across the collector-base SCR by drift. The electric field is very high so the electrons reach their saturated velocity almost immediately. An expression for that is:

$$\tau_{BC} = \frac{W_{BC}}{2v_{sat}} \tag{1.9}$$

where W_{BC} is the collector-base depletion region width. An increased base doping reduces the space-charge width thus improving τ_{BC}.

While the first-order approximation in Eq. 1.3 is valid at DC and low frequency, problems arise while investigating the dynamic behavior of transistors in fast switching operation. Starting from around 0.5 f_T, a certain delay is observed during switch-on and switch-off.

Winkel [136] observed a close similarity between the differential equations of currents inside the transistors and the flow of a signal in a transmission line. In particular, when the carriers' lifetime and the electric field in the transistor are constant, the analogy will be a uniform transmission line. The changes in the "phase of the signal", in our analogy, are, in the time domain, describable by constant delays. The current gain and the transconductance become:

$$\beta_F' = \frac{\beta_0}{1 + j\dfrac{f}{f_T}\beta_0} \cdot \exp\left(-j2\pi f \tau_1\right) \qquad g_m' = g_m \exp\left(-j2\pi f(\tau_1 + \tau_2)\right) \tag{1.10}$$

These equations are applied to transistor's compact models to describe *non quasi-static* (NQS) effects.

Maximum Oscillation Frequency f_{max} is tightly linked to the definition of *maximum available power gain* (MAG), G_{ma}, and *maximum stable power gain* (MSG), G_{ms}, of a transistor [45]. The MAG can be expressed in terms of S-parameters as:

$$G_{ma} = \left|\frac{s_{21}}{s_{12}}\right| \left(k \pm \sqrt{k^2 - 1}\right) \tag{1.11}$$

where k is the so-called *stability factor*: when $k < 1$, the MAG is not defined (complex), and in particular, when $k < 0$, oscillations will occur. When $k = 1$, on the other hand, $G_{ma} = G_{ms}$, and the device is just stabilized. We define the frequency at which the MAG and MSG become one, f_{max}. This frequency is also the frequency at which Mason's unilateral gain U becomes unity, i.e.:

$$\left|U(f)\right|_{f=f_{MAX}} = 1 \tag{1.12}$$

In a two-port active device, if the reverse signal flow is much smaller than the forward flow, it can be approximated by 0: in essence, this figure of merit states whether this simplification affects the device accuracy. It can be also shown that U is inversely proportional to the square of the frequency, hence we can write:

$$|U(f)| \propto f^{-2} \Longrightarrow |U(f)| = \left(\frac{f_{max}}{f}\right)^2 \tag{1.13}$$

Mason's parameter was found to be invariant with respect to any linear, lossless and reciprocal embedding, and therefore represents a figure of merit to compare any two-port active device. It can be calculated from Y-parameters as:

$$U(f) = \frac{\left|y_{21} - y_{12}\right|^2}{4\left(\mathrm{Re}\left(y_{11}\right)\mathrm{Re}\left(y_{22}\right) - \mathrm{Re}\left(y_{12}\right)\mathrm{Re}\left(y_{21}\right)\right)} \tag{1.14}$$

From this equation, we conclude that, if U is greater than one, the device under test is active, otherwise, we conclude it is passive. f_{max} is a characteristic of the device, and can be defined either as the highest frequency at which an oscillator made with a single active device and embedded in a passive network can establish an oscillating behavior (thus explaining its name), or as the maximum frequency for which a transistor, embedded in a passive network, can amplify power.

For bipolar transistor we can approximate f_{max} by the following formula [72]:

$$f_{max} \approx \sqrt{\frac{f_T}{8\pi\left[\left(R_{Bx} + R_{Bi}\right)C_{BCi} + R_{Bx}C_{BCx}\right]}} \approx \sqrt{\frac{f_T}{8\pi R_B C_{BC}}} \tag{1.15}$$

where R_B is the base resistance and C_{BC} the base-collector capacitance, and in the second approximation they are not separated into intrinsic (subscript "$_i$") and extrinsic (subscript "$_x$") contributions (see Fig. 1.9).

As we can clearly see from Eq. 1.15, the great decrease of the base resistance in HBTs mirrors on a higher f_{max}, which is made even bigger by the overall increase of f_T as well, due to the decrease of τ_E and τ_{BC} by fabrication.

In conclusion, while the frequency f_T gives a measure of the speed of switching circuits such as dividers, f_{max} represents a speed metric for circuits such as amplifiers and oscillators, and help understand the ability of an active device to absorb power from a source (port 1), amplify it and deliver it efficiently to a second terminal (port 2) [90, 108]. Other relevant figures of merit for evaluating potential applications of HBTs include the minimum noise figure, the linearity, and the gain of a transistor, but they will not be treated here.

1.2.3 Silicon-Germanium HBTs

Existing semiconductors can be divided into two categories: elemental and compound semiconductors. Within the first group are germanium and silicon. Because of its diamond-like lattice structure and some interesting properties, like high electron mobility, germanium was one of

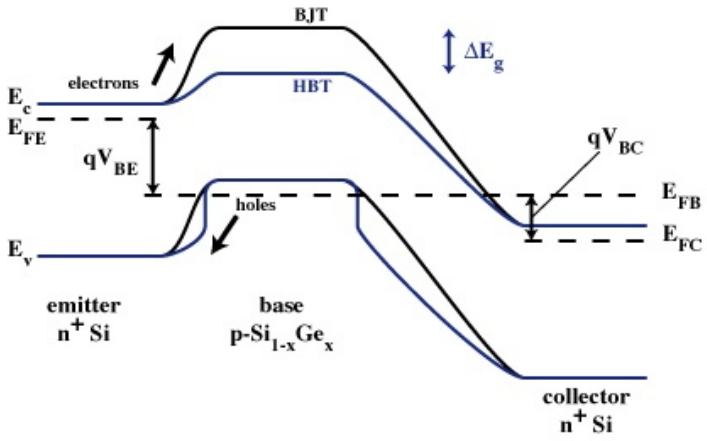


Figure 1.7: Different displacement of the conduction band in BJT and SiGe HBT (after [74]).

the first semiconductors applied for early transistors. Nevertheless, silicon has a bigger bandgap (1.12 eV versus 0.67 eV), showing less intrinsic-carriers growth with temperature, so it is nowadays mainly used in the semiconductor industry. Yet, germanium has been widely applied in the last decades in compounds.

Compound semiconductors are materials exploited for bandgap engineering in devices like HBTs. Those can be alloys of elemental materials, like SiGe, or alloys formed by elements from two different groups of the periodic table, like III-V compounds (GaN, GaAs and InP). HBTs using indium phosphide (InP) as main component combine it with other materials which are lattice-matched to it [111, 123]. Several ternary alloys can be used as the base, such as $In_{0.52}Al_{0.48}As$ (indium aluminium arsenide, InAlAs), $In_{0.53}Ga_{0.47}As$ (indium gallium arsenide, InGaAs), $GaAs_xSb_{1-x}$ (gallium arsenide antimony, GaAsSb). InGaAs has E_g = 0.75 eV, GaAsSb has (on average) E_g = 0.72 eV [18], while InP has a bandgap E_g = 1.35 eV. This remarkable bandgap difference provides an orders-of-magnitude higher current gain than any other solution. In the following, however, we will focus on silicon-based HBTs, in particular bipolar CMOS only, for their best integration and other reasons listed in the dedicated paragraph below.

In a SiGe HBT, the base p-type semiconductor is composed of an alloy of silicon and germanium $Si_{1-x}Ge_x$, x being the atomic percentage of germanium in the alloy. The electron affinity of SiGe is similar to that of Si, so that the conduction band discontinuity is small, or equivalently we can substitute in Eq. 1.1 ΔE_g with ΔE_v, the difference between the levels of the valence bands. SiGe HBTs have achieved an average percentage of germanium of $x = 0.2$, resulting in ΔE_v = 200 meV [111].

From the carriers point of view, the Fermi level in the p-type base is closer to vacuum in HBTs than it is in homojunction BJTs. This means that the displacement of conduction bands when the two regions are in contact is less for heterojunctions, hence the barrier for electrons is lower than in the EBJ of BJTs (Fig. 1.7) [72].

When the EBJ is created, dislocations may form because the lattice constant of the alloy is over 4% larger than that of Si, but as seen in [68], if a critical strained-layer thickness is respected by limiting the dose of germanium, the misfit adjusts itself elastically and no defect appears (pseudomorphic growth). The bandgap of SiGe is usually several tenths of an electron-volt smaller than that of Si, but pseudomorphic growth makes it even smaller. In fact, SiGe is now considerably strained and this has a beneficial effect on the transistor properties by further reducing the bandgap of SiGe.

If HBTs are built at low temperatures, dislocations due to incoherently strained layers tend to be less, using the same germanium fraction. This is why for the fabrication process, molecular beam epitaxy (MBE), which allows pseudomorphic growth at relatively low temperatures, is preferred. Other deposition methods are chemical vapour deposition (CVD) and rapid thermal chemical vapour deposition (RTCVD), which allows selective epitaxial growth (SEG) on patterned

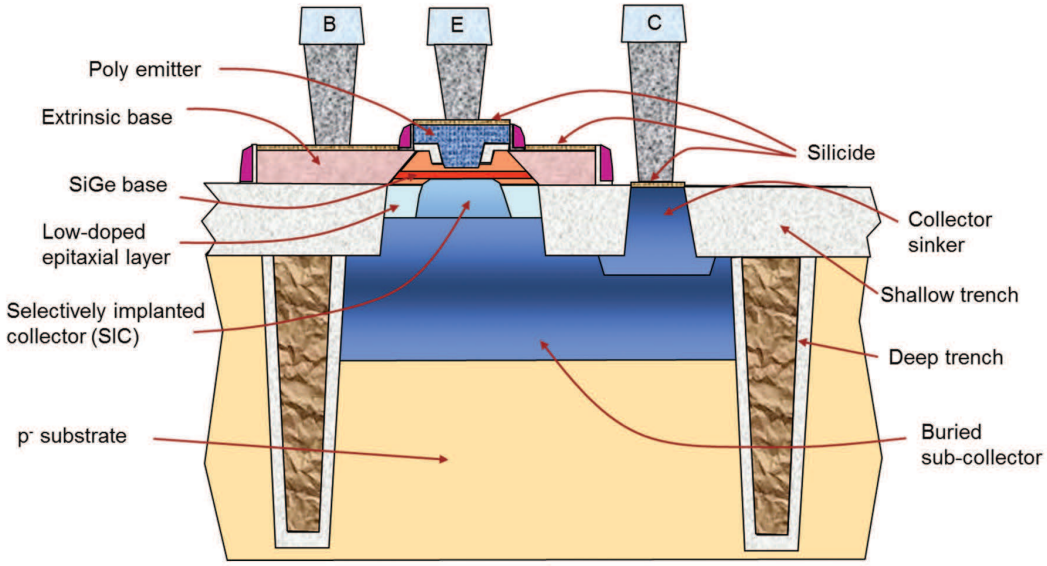


Figure 1.8: Schematic cross-section of high-speed SiGe HBT device architecture (after [90]).

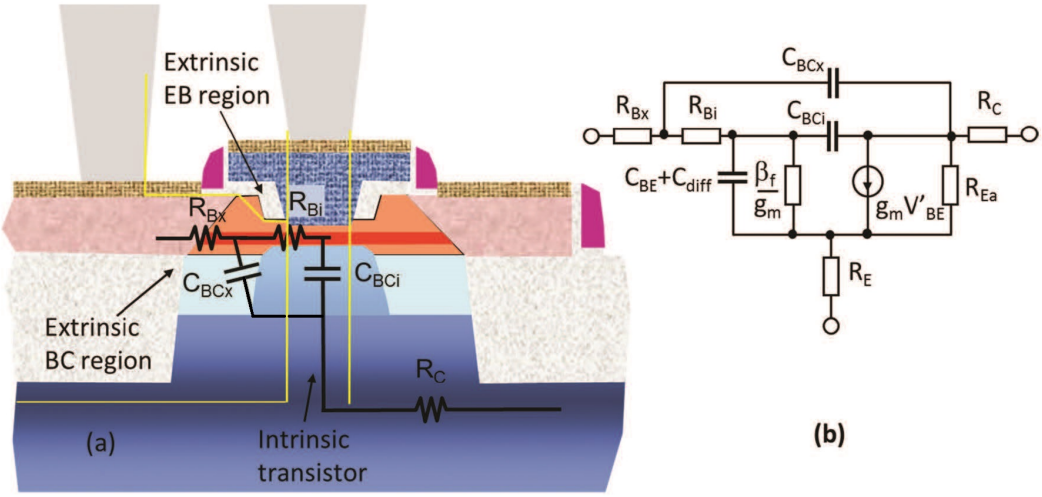


Figure 1.9: Device cross section with parasitic resistances and capacitances associated with different device regions (a) and a corresponding small signal equivalent circuit (b) (after [90]).

regions. In Fig. 1.8 an example of SiGe HBT structure is presented, with associated parasitic resistances/capacitances composing a small-signal transistor model, in Fig. 1.9 (the model is simplified, hence some elements, such as parasitic oxide capacitance between base and collector contact, or STI oxide capacitance between poly base and collector, are excluded). SiGe HBT processes a huge advantage over the III-V HBTs in that the fabrication is more mature and SiGe HBT can be fabricated with the existing CMOS technology with only few more steps needed.

1.2.4 Bipolar CMOS

BiCMOS technology (Bipolar Complementary Metal Oxide Semiconductor) integrates in the same device both a classical CMOS technology and a bipolar transistor, taking advantage of the properties of both technologies [2]:

- low power consumption (from CMOS);
- very good analogue amplifier (the CMOS gives high input impedance while the bipolar makes output impedance low);

- low variability in electrical parameters to temperature and process variations;
- high current gain (from bipolar), making it suitable for long-lasting remote applications, for example;
- higher packaging density for logic (from CMOS);
- in a series configuration, its total capacitance is low (almost as much as the bipolar), this leading to better frequency performance as a broadband amplifier or high switching speed in digital applications;
- good fan-out, i.e. it can drive high capacitance load with reduced cycle time (no buffers needed, unlike CMOS);
- latch-up invulnerability;

The main drawback is their high complexity in fabrication, thus high costs, with respect to pure CMOS technology. However, advancements in Si technology, growing demand for reliable mmW ICs, lower manufacture costs and much easier integration have made SiGe HBTs highly competitive in mmW applications.

1.2.5 Technology Under Analysis

In this work, the 55 nm SiGe:C BiCMOS (BiCMOS055, or B55 for short) by STMicroelectronics will be studied [16, 17]. ST's BiCMOS combines a CMOS and a npn SiGe HBT, the architecture of which has been developed through generations from the early single-polysilicon quasi self-aligned architectures to modern double-poly fully self-aligned (DPSA) – whose technology is used for the B55 too. Owing to their digital density 5 times higher than previous 130 nm technology, B55 well serves optical, wireless and high-performance analogue applications.

B55 is based on a 55 nm triple gate CMOS platform (55 nm being the average distance between identical features in an array of memory cells made with this technology), featuring both Low Power and General Purpose CMOS.

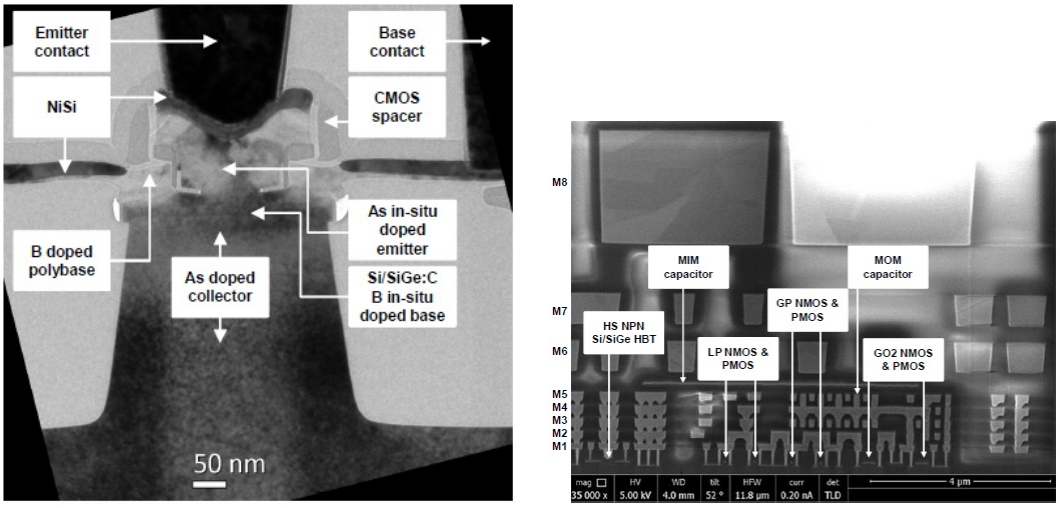
The SiGe HBT has to provide minimum access resistances to the emitter, base and collector, and minimum base-collector and base-emitter capacitances as well. It features the double-poly self-aligned architecture with selective epitaxial growth (DPSA-SEG) of the boron-doped SiGe:C base (C hinders boron diffusion) (Fig. 1.8a).

The double-poly architecture provides access from the contact regions to the emitter and intrinsic base regions by poly-Si layers which dielectrically isolate against the surrounding transistor regions: R_E, R_{Bx}, C_{BE}, C_{BCx} are kept small.

While this practice is quite common, what differentiates the HBT fabrication is the growth of the base. The SEG compared to non-selective epitaxial growth is the most attractive process for its simplicity (since only one lithographic step is needed) and its effectiveness in the quality of self-alignment of the emitter with the base. It is inserted in the fabrication process between gate poly-Si deposition and gate patterning of CMOS to reduce the total amount of thermal energy transferred to the CMOS during elevated temperature operations. However, improvement compared to the current performance level are hardly possible with this approach, due to the impossibility to cut the extrinsic base resistance R_{Bx} [8].

ST's SiGe HBT (Fig. 1.10a) comes in three collector flavours, with different $f_T \times BV_{CEO}$ trade-offs: High Speed (HS), Medium Voltage (MV) and High Voltage (HV). The back-end (made of 8 copper layers and an aluminium cap, Fig. 1.10b) will be extensively considered in the following; it is fully compatible with CMOS and provides enhanced mmW performance [18].

f_T and f_{max} performances are obtained thanks to the higher collector current densities and reduced parasitic resistances and capacitances allowed by the vertical and lateral scaling of transistors. In Fig. 1.11, we can observe such an improvement from older technologies notably at



(a) TEM cross-section of a $0.1 \times 4.9\,\mu m^2$ HS SiGe HBT (zoom in on emitter and base regions)

(b) SEM cross-section showing main devices and back-end up to metal 8

Figure 1.10: BiCMOS055 by STMicroelectronics (after [16]).

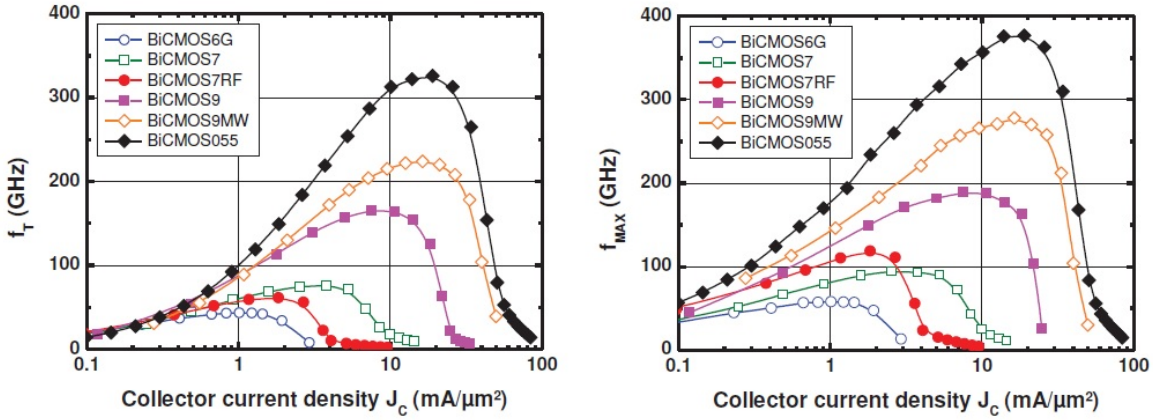


Figure 1.11: Evolution of f_T and f_{max} vs. collector current density in STMicroelectronics SiGe BiCMOS technologies (after [17]).

relatively high collector current density, with HF performance standing out for B55. Then the performance is degraded by heavy injection effects. By increasing f_T and f_{max}, the collector-emitter breakdown voltage consequent reduction points out the need for a trade-off. However, the $f_T \times BV_{CEO}$ of B55 is the highest of all ST's BiCMOS.

Whereas early 90 nm BiCMOS technologies featured HBTs with HF performance that reached $(f_T, f_{max}) = (130, 100)$ GHz [57], measured results for the SiGe HS HBT in B55 prove a better HF behaviour, the couple of values being $(f_T, f_{max}) = (326, 376)$ GHz.

Advancements on the BiCMOSB55 process led STMicroelectronics to develop the next bipolar CMOS technology: BiCMOS55X. The optimization of the vertical profile including the thermal budget and the fully implanted collector development showed a very good potential (450 GHz f_T was demonstrated in [42]). The new emitter/base architecture called EXBIC [122] is certainly the most challenging part of these developments. Such an architecture aims at the realization of the targeted 600 GHz f_{max} performance.

Chapter 2

Measurement and Calibration Basics

Contents

THE PROGRESS of electronics over the course of the years has made possible to build complex circuits made of devices operating at very high frequencies, and this trend is expected to continue. The growing complexity of high frequency integrated circuits relies on the quantification of several nano-electronic device properties that need to be precisely modelled and characterized.

Indeed, it is fundamental to ensure that a device for any millimeter-wave application behaves as expected in order to provide circuit design engineers advanced and accurate libraries (design kits) for CAD platforms. The libraries implement lumped electrical circuit-based models, whose parameters have to be extracted from trustworthy millimeter wave frequency measurements and verified and elaborated by simulations, either of passive or active devices (in this case, through compact models, depicting the underlying physics, based on semiconductor, electromagnetical and thermal equations). For instance, if a designer employ a transistor to work at a certain frequency in his/her system, he or she expects a certain output power from the model, which therefore needs to rely on precisely characterized and well-calibrated measurements. Therefore, much of the semiconductor HF characterization deals with the mathematical removal of test fixtures and on-wafer parasitic elements.

In this chapter, we will describe how small-signal measurements are performed and learn that, in fact, the raw measurement data themselves do not describe at all the intrinsic device behavior, and need for additional data processing steps.

The most wide-spread measurement system, the vector network analyser, will be described and the signal quantities will be rigorously defined, before plunging into an overview of the assembly parts of network analysers, with their functions and components accurately portrayed. The notion of measurement error terms will let us introduce some of the most common calibration routines, which effectively provide a solution for correcting the errors. The challenges of going up in frequency and perform measurements, the dedicated setup and adjustments of calibrated data, and, finally, the used measurement configuration and instrumentation, will conclude the chapter.

2.1 Vector Network Analyser

2.1.1 Signals and Scattering Parameters

Vector network analysers (VNAs) are measurement systems which are used to analyse circuits from small components (transistors, filters, amplifiers, etc...) to more complex modules by comparing the real and imaginary part of an incident signal (the one generated by the VNA) with the transmitted signal (the one passing through the system and measured on another side) or the reflected signal (the one reflected by the input of the system).

Those RF signals are defined by convention as "a" for the incident wave and "b" for the reflected (or transmitted) wave. "a" and "b" are the so-called *"power waves"* and their squared absolute values (i.e. "$|a|^2$" and "$|b|^2$") represent the true incident and reflected powers, respectively. In a general way, they are defined as:

$$a = \frac{\mathrm{V} + \mathrm{Z_0 I}}{2\sqrt{\mathrm{Re(Z_0)}}} = \frac{\mathrm{V^+}}{\sqrt{\mathrm{Re(Z_0)}}} \tag{2.1}$$

$$b = \frac{\mathrm{V} - \mathrm{Z_0^* I}}{2\sqrt{\mathrm{Re(Z_0)}}} = \frac{\mathrm{V^-}}{\sqrt{\mathrm{Re(Z_0)}}} \tag{2.2}$$

where $\mathrm{Z_0}$ is the characteristic impedance of the line and $\mathrm{V^+}$ and $\mathrm{V^-}$ are the incident and reflected voltage wave amplitudes, respectively. We can thus observe that the power waves carry the same information as the voltages. For a lossless line ($\mathrm{Z_0}$ positive and real), the reflection coefficient at port i is:

$$\Gamma_i \triangleq \frac{\mathrm{V_i^-}}{\mathrm{V_i^+}} = \frac{b_i}{a_i} = \frac{\mathrm{Z}_i - \mathrm{Z_0}}{\mathrm{Z}_i + \mathrm{Z_0}} \tag{2.3}$$

where Z_i is the impedance at port i.

VNAs can perform several measurements of these quantities and retrieve fundamental information about the system under test, such as the *scattering parameters* (S-parameters). These quantities describe the electrical behavior of a linear devices and are defined as ratios of the reflected (or transmitted) waves to the incident ones. Some VNAs can be mounted with up to 48 ports [105]. In general, for a n-port network:

$$\begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \begin{bmatrix} s_{11} & \cdots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nn} \end{bmatrix} \cdot \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \tag{2.4}$$

where b_i and a_i are the incident/reflected waves, respectively, for each port i, which is terminated to a specific Z_{0i} impedance. For a 2-port network (Fig. 2.1), however, the previous formula can be reduced to the following system:

$$\begin{aligned} b_1 &= s_{11} a_1 + s_{12} a_2 \\ b_2 &= s_{21} a_1 + s_{22} a_2 \end{aligned} \tag{2.5}$$

so it's easy to define:

$$\begin{aligned} s_{11} &= \left.\frac{b_1}{a_1}\right|_{a_2=0} & s_{21} &= \left.\frac{b_2}{a_1}\right|_{a_2=0} \\ s_{12} &= \left.\frac{b_1}{a_2}\right|_{a_1=0} & s_{22} &= \left.\frac{b_2}{a_2}\right|_{a_1=0} \end{aligned} \tag{2.6}$$

Scattering parameters are particularly suited for active devices under test (DUTs) such as transistors, considering that, for defining other electrical parameters (such as Y, Z and H), it would be

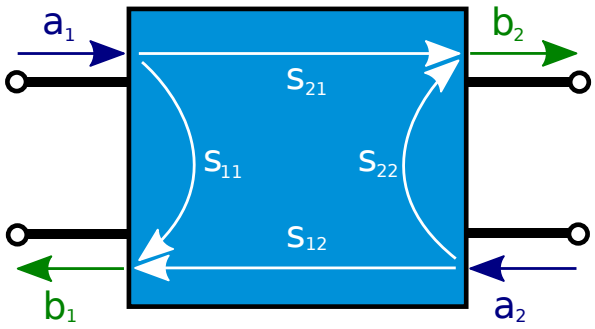


Figure 2.1: A 2-port network where signals and S-parameters are shown.

necessary to effectively create short and open circuits for measuring them, which is extremely hard and imprecise, particularly at high frequencies (tuning stubs would be needed) [11].

Flow charts are useful schematic descriptions of networks where signals and parameters are presented in blocks so that the relation between them can be easily explicated by following the block-to-block connections. As way of example, Fig. 2.2 shows an equivalent flow chart indicating all coefficients appearing in a 2-port network when a real voltage source is applied at port 1 (voltage V_g, impedance $Z_g \neq Z_0$) and a generic load at port 2 ($Z_L \neq Z_0$).

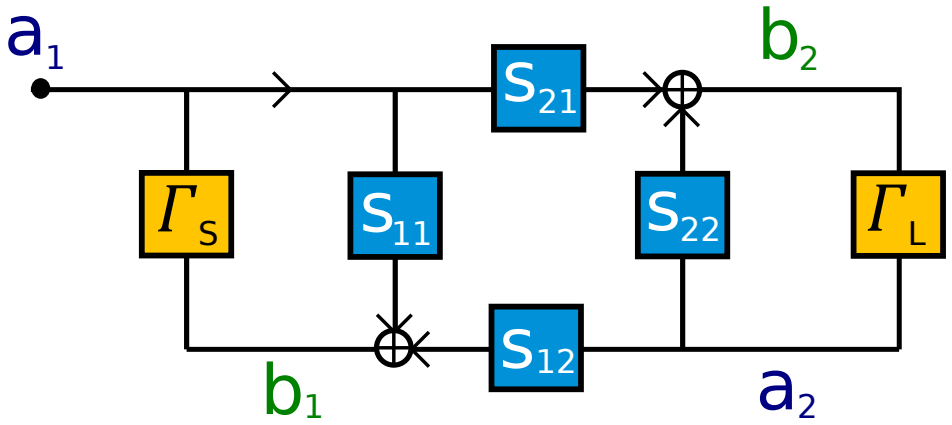


Figure 2.2: Flow chart representing all reflection and transmission coefficients for a 2-port network when applying a voltage generator at port 1 at a given load at port 2.

In this case, a_1 can be written as:

$$a_1 = \frac{V_1 + Z_0 I_1}{2\sqrt{Z_0}} = V_g \frac{\sqrt{Z_0}}{Z_g + Z_0} \tag{2.7}$$

where the second equation comes from a simple voltage divider at the entrance of the DUT, $V_1 = V_g \dfrac{Z_0}{Z_g + Z_0}$, supposing that the impedance of the network is the characteristic impedance. From the previous diagram, the input reflection coefficient for a generic load is:

$$s'_{11} = \left.\frac{b_1}{a_1}\right|_{\Gamma_L \neq 0} = \frac{a_1 s_{11} + a_1 c}{a_1} \tag{2.8}$$

where c takes into account every contribution from the input of the DUT on, and can is calculated as:

$$c = \Gamma_L \left(\frac{c}{s_{12}} s_{22} + a_1 s_{21} \right) s_{12} \tag{2.9}$$

By rearranging this equation and putting all together, the previous equation yields:

$$s'_{11} = s_{11} + \frac{s_{21}\, s_{12}\, \Gamma_L}{1 - s_{22}\, \Gamma_L} \tag{2.10}$$

This equation shows that, when excited by an external stimulus, the reflected wave at the entrance of the stimulated system depends on all the scattering parameters of the network, as well as, by the reflection of the signal, on the load. It is only when the load is perfectly matched to the characteristic impedance of the network ($Z_L = Z_0$, or, equivalently, $\Gamma_L = 0$), that s_{11} equals the reflection coefficient at port 1, s'_{11}. Also, one can realize that by measuring four well-known loads, all the S-parameters of the network can be found.

2.1.2 Architecture of the VNA

S-parameters are complex numbers and in microwave devices, they can be employed to calculate some important qualities, or *figures of merit*, such as gain, losses, reflection and amplification, over a certain range of frequency. Vector network analysers are able to evaluate both magnitude and phase of the S-parameters of a DUT (single or multi-port), by sweeping up to THz frequencies (with the use of dedicated extenders, as it will be discussed later) and by controlling the injected power at each port. A *scalar network analyser* (SNA), which compared to VNA can only measure magnitudes of signals, is less pricey and can be used for similar purposes. However, some of the benefits of a VNA over a SNA are [47]:

- full system error correction (for systematic errors);
- complex parameters can be translated to the time domain;
- de-embedding/embedding capabilities;
- Smith chart drawing.

In this manuscript, only VNAs are used as measuring tools.

In Fig. 2.3, a simplistic block diagram of the main components of a VNA is shown [47]. It can be grouped into four blocks according to its functions [116]:

- a *generator stage*, to output a stimulus in two different modes, as power and frequency sweep;
- a *signal separation stage* (or *test set*), which separates the forward and reverse waves. A *power splitter* and a *directional element* are part of it: the former provides the same values for a_1 and a'_1, that are proportional to the generated signal ($a''_1 = \alpha a'_1 = \alpha a_1$, α being an arbitrary proportionality constant), while the latter (a coupler, at high frequency), is in charge of separating the wave directed to the DUT from the one reflected;
- the *test port*s, which is the physical interface between the VNA and the DUT;
- a *receiver and analyser module*.

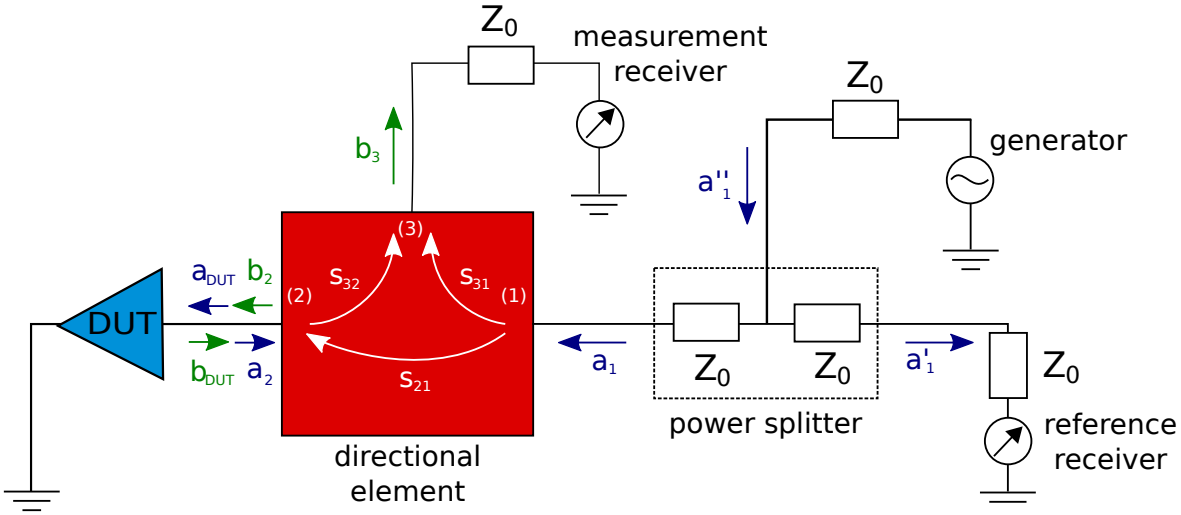


Figure 2.3: Schematic representation of an n-port VNA connected to a 1-port DUT.

Signal Generator The first stage deals with the signal generation. Engineers have many degrees of freedom for the choice of the shape of signal in modern VNAs; which can be sinusoidal, as well as pulsed, squared, modulated, etc... For this purpose, a *voltage-controlled oscillator* (VCO) such as a crystal oscillator plus a feedback, provided by a *phase locked loop* (PLL), generate a signal sweeping in a desired frequency range. In addition to that, an *automatic loop control* (ALC) stage ensures to deliver a controllable power to the DUT and a proper level for the operation in the case of an active device.

Directional Element Next, both the generated and received signals have to be sampled in order to compute the S-parameters. Ideally, the VNA behaves in such a way that the reflection coefficient of the DUT is given by the ratio of the measured wave b_3 and the incident wave a_1 (Fig. 2.3). Indeed, the directional element adds its contribution to the *measured value* $\mathrm{M} = b_3/a_1'$, so that $\mathrm{M} = \mathrm{R}\Gamma_{\mathrm{DUT}}$, where $\Gamma_{\mathrm{DUT}} = b_{\mathrm{DUT}}/a_{\mathrm{DUT}} = a_2/b_2$. R is called *reflection tracking*, and is defined as $\mathrm{R} = s_{21}s_{32}$: it takes into account the attenuation due to the passage of the signal through the directional element. We are assuming here that a_1 is constant and identical to the one used as a reference, which is achieved by the use of the splitter. Yet, we are not considering some non-ideal effects which affect the measurement. First of all, $s_{31} \neq 0$, which means that a certain quantity of the input signal deviate from the path to the DUT and adds up to the measured power wave. We can thus write:

$$\mathrm{M} = \mathrm{R}\left(\frac{s_{31}}{s_{32}s_{21}} + \Gamma_{\mathrm{DUT}}\right) = \mathrm{R}\left(\mathrm{D} + \Gamma_{\mathrm{DUT}}\right) = \mathrm{R}\,\Gamma_{\mathrm{DUT}}\,(x+1) \tag{2.11}$$

We introduce the quantity D, called *directivity*. $x = \mathrm{D}/\Gamma_{\mathrm{DUT}}$ is a relative deviation due to a systematic error given by s_{31} and requires correction. Also, multiple reflections at port 2 of the coupler should be considered. If we stop at the second reflection, after which the reflected power is negligible, we get:

$$a_2 = b_2\,\Gamma_{\mathrm{DUT}}\left(1 + \mathrm{S}\,\Gamma_{\mathrm{DUT}}\right) = b_2\,\Gamma_{\mathrm{DUT}}\,(1+y) \tag{2.12}$$

where S is the reflection coefficient of the test port (s_{22}), also called *test-port match*, and it contributes as a relative error $y = \mathrm{S}\,\Gamma_{\mathrm{DUT}}$. So by summing up all the results, the previous formulas yield:

$$\mathrm{M} = \mathrm{R}\left[\mathrm{D} + \Gamma_{\mathrm{DUT}}\left(1 + \mathrm{S}\,\Gamma_{\mathrm{DUT}}\right)\right] \tag{2.13}$$

The directivity and the test-port match limit the measurement accuracy by adding uncertainty to the measurement. If $|\Gamma_{\mathrm{DUT}}|$ is small, the error given by D will limit the measurement, while if $|\Gamma_{\mathrm{DUT}}|$ is large, S will be the bound. R is independent of $|\Gamma_{\mathrm{DUT}}|$, whereas it is not the case for D and S.

Test Ports The connection between the VNA and the device are the so-called "*ports*" and the majority of VNAs comes with 2 or 4 of them. Often times, they are coaxial sockets located in the front side of the network analyser, allowing to employ coaxial cables. The use of coaxial cables rather than rectangular waveguides to carry the signals is made, of course, for simplicity of handling thanks to their flexibility, and also because of the strong advantage to produce a pure TEM wave, thus presenting no low frequency cutoff, and withstand a broad spectrum of frequency from DC to tenths of gigahertz. To connect and contact them to the device ports, additional elements are needed, such as tees and *high frequency* (also called *radio frequency*, RF) *probes*. Non-idealities in the signal separation stage and in the test ports, however, affect the quality of the measurement and some procedures grouped under the name of "calibration", which are at the core of this work, need to be performed to "clean off" the raw measurements.

Receiver Finally, the receiver works according to the heterodyne principle; the wave coming from the test set can be written as:

$$x_{RF}(t) = A_{RF} \exp\left[j\left(2\pi f_{RF}\, t + \varphi_{RF}\right)\right] \tag{2.14}$$

This signal is fed into a mixer together with a signal coming from a local oscillator. The frequency of the input signal f_{RF} is modified by using the tunable f_{LO}, the frequency of the oscillator, so to match the desired intermediate f_{IF}, the frequency selected by the IF filter, according to $f_{IF} = |f_{RF} - f_{LO}|$: the anti-aliasing filter thus made removes any broadband noise. Then the resulting signal, which can be expressed as:

$$x_{IF}(t) = A_{IF} \cos\left(2\pi f_{IF}\, t + \varphi_{IF}\right) \tag{2.15}$$

where $A_{IF} = \dfrac{A_{RF} A_{LO}}{2}$ and $\varphi_{IF} = \varphi_{RF}$, is demodulated by a signal generated by a *numerically controlled oscillator* (NCO). This phase is called *synchronous detection* (or *synchronous demodulation*), and consists on a down conversion to DC in an in-phase/quadrature sense. The NCO-generated signal, defined by:

$$x_{NCO}(t) = A_{NCO} \cos\left(2\pi f_{NCO}\, t\right) \tag{2.16}$$

and its 90 degrees-shifted form, $x'_{NCO}(t) = A_{NCO} \sin\left(2\pi f_{NCO}\, t\right)$, are fed into two separate multipliers with $x_{IF}(t)$. If we select $f_{NCO} = f_{IF}$, the multipliers output the following signals:

$$\begin{aligned} x_Q(t) &= \frac{A_{IF} A_{NCO}}{2} \left[\cos\left(\varphi_{IF}\right) + \cos\left(4\pi f_{NCO}\, t + \varphi_{IF}\right)\right] \\ x_I(t) &= \frac{A_{IF} A_{NCO}}{2} \left[\sin\left(\varphi_{IF}\right) - \sin\left(4\pi f_{NCO}\, t + \varphi_{IF}\right)\right] \end{aligned} \tag{2.17}$$

Finally, a low-pass filter suppresses all the components at $f \neq 0$ and keeps the DC terms:

$$\begin{aligned} x_Q &= \frac{A_{IF} A_{NCO}}{2} \cos\left(\varphi_{IF}\right) \\ x_I &= \frac{A_{IF} A_{NCO}}{2} \sin\left(\varphi_{IF}\right) \end{aligned} \tag{2.18}$$

from which we can derive φ_{IF} and A_{IF}, hence φ_{RF} and A_{RF}. Then the resulting signals are fed into an ADC, which is now operating at DC: that allows a simpler clocking structure.

2.1.3 Measurement Errors and Calibration Techniques

Every raw measurement result is affected by errors introduced by the measuring setup. These errors may be of various kinds [85, 7, 96]:

- *Random errors* are caused by a lack of repeatability in the output of the measuring system, they are statistically describable but no systematic correction is possible. They are due to instrument noise, repeatability errors (e.g. different probes position during different measurements, EM interference, etc...). Noise is an electrical perturbation due to the components of the VNA, in particular the local oscillator in the receiver (phase noise). This kind of error can be made negligible by inputting a higher power level or by reducing the IF filter bandwidth. Random errors can be corrected only if their statistical description has zero average, by averaging multiple measurements.

- Errors due to the *non-linearity* of the DUTs. When dealing with non-linear devices (such as bipolar transistors), spurious harmonics may be generated when the input power is high. The theoretical linearity of power gain is gradually lost as the input power rises. After a

certain point the output signal goes into compression as the gain flattens (1-dB compression point). The device's response becomes non-linear and produces distortions. The input must be reduced, so it's about a trade-off between non-linearity and random errors.

- *Drift errors* are due to the characteristics of the instrument that change with time (thermal dilatation of cables, resistances changing at the contact level...). A controlled temperature during the measurement process avoids major thermal drifts. A re-calibration may solve this issue.

- *Systematic errors* are due to imperfections of the instrument and connections; they are consistent and repeatable (i.e. predictable) and do not change over time. A systematic correction can be applied by knowing the errors and the measurements as vector quantities. This correction is called *calibration*, and is regularly needed to avoid the aforementioned drift errors. Also, using different cables and adapters will change the properties of the system, thus requiring a new calibration.

Hence, calibration is the process through which we can determine the above-mentioned *error terms* (also known as *correction data*) of an error model, a mathematical representation of the systematic errors' contribution in the measuring system; these non-idealities include power loss in the waveguide section, extenders and probes, reflections due to imperfect matching between various test fixtures, leakage and directivity errors in the VNA [146]. The issues related to network analyzer characterization were first tackled in the 1970s [5]. The goal is to obtain the physical characteristics (i.e. the S-parameters) at a well-defined *reference plane*, either at probe tips for commercial substrates or closer to the DUT for on-wafer calibration. The error terms are found by connecting the VNA to a certain number of *calibration standards*, which are networks with known properties. Once the error terms calculated, more accurate DUT's S-parameters can be retrieved from raw S-parameters: in other words, the properties of the VNA and the test assembly are excluded from DUT measurements and the corrected measurement represents a more accurate estimate of the S-parameters of the transistor. How many and which standards should be used depends on the calibration and the selected routine.

3-Term Error Model For a 1-port network, a simple 3-term error model is used; these terms can be found by a SOL (*Short-Open-Load*) calibration technique. Let us consider again the diagram of Fig. 2.3. We can group all the non-ideal contributions we previously described in a 3-term error model. Anything between the generator/receiver and the DUT (the test set but also the cables, adapters, etc...) is included into a 2-port network with coefficients e_{11}, e_{21}, e_{12}, e_{22}, but can be further simplified to just three terms since there is no transmission inside the DUT, as we see in Fig. 2.4:

- $e_{11} = e_D$ is the directivity. It represents the part of the signal going from the generator directly into the receiver (and never reaching the DUT), due to the non-ideal coupler;
- $e_{22} = e_S$ is the source match. The signal reflected by the DUT reaches port 2 and is reflected back to the DUT, where it combines with the incident signal. It is due to the imperfect output impedance of the VNA;
- $e_{12}e_{21} = e_R$ is the reflection tracking (no need to differentiate between the single transmission coefficients). It is the non-unitary frequency response of the measuring system, the signal path inside the test channels mainly.

The measured quantity M essentially is the portion of the generated signal which is transmitted to the receiver. Analogously to Eq. 2.10, it can thus be written:

$$\mathrm{M} \equiv \frac{b_3}{a_1} = e_D + \frac{e_R\,\Gamma_{DUT}}{1 - e_S\,\Gamma_{DUT}} \tag{2.19}$$

In this formula, Γ_{DUT} is unknown, as well as all the error terms. Theoretically, we could use any kind of calibration standards, provided that the reflection coefficients are well-known. However, it is important to choose standards with properties as different as possible from one another: that

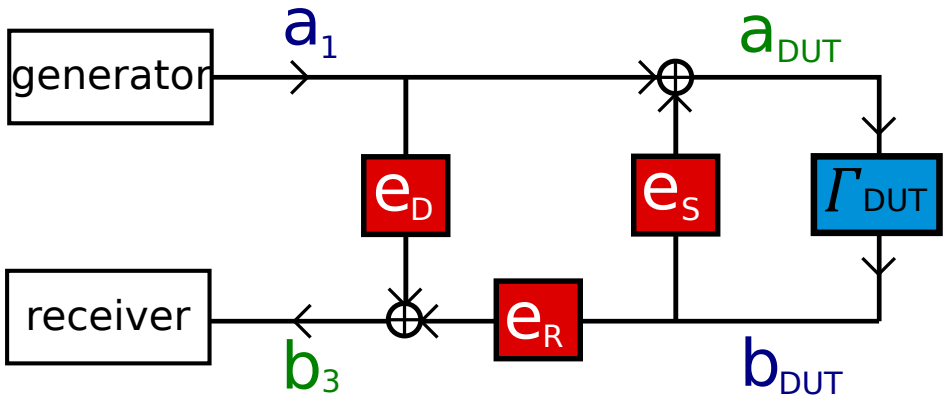


Figure 2.4: 3-term error model.

will maximise the dynamic range of the VNA and give no ambiguity on the measured quantities. If we use a short, a 50Ω load and an open as our DUTs and we measure them, we get: M_S ($\Gamma_{DUT} = \Gamma_S$), M_O ($\Gamma_{DUT} = \Gamma_O$), M_L ($\Gamma_{DUT} = \Gamma_L = 0$). By reversing the previous formula, all the error terms are yielded and this expression follows:

$$\Gamma_{DUT} = \frac{M_{DUT} - e_D}{e_R + (M_{DUT} - e_D)\, e_S} \tag{2.20}$$

from which we can finally compute Γ_{DUT}.

12-Term Error Model The error model for a 2-port network can be developed in the same way as the 1-port [46] (Fig. 2.5). Older and cheaper VNAs have three couplers to direct the signal: one is for directing the generated signal to the reference receiver, the others direct the reflected/transmitted signal to the measurement receiver, one for each port (port 1 and port 2). A switch is used to direct the signal to port 1 or 2, while the other port is terminated to a 50Ω impedance. If the characteristics of the switch change by changing its position, then we need to distinguish two cases: one in which the generated signal a_0 is sent to port 1, while port 2 is terminated on the reference impedance so that there is no a_3 signal; one in which the generated a'_3 goes to port 2. All the imperfections due to the switch are taken into account by the error model.

At the input port we can find again the three error terms we have already encountered in the 1-port network, i.e. forward and reverse directivity (e_{DF}, e_{DR}), source match (e_{SF}, e_{SR}), reflection tracking (e_{RF}, e_{RR}). However, the signal passing through the output port will generate new sources of error:

- e_{LF} and e_{LR} represent the *load match*. The signal is transmitted by the DUT and partly reflected by the output port back to the DUT. This wave will be measured by the input port.
- e_{TF} and e_{TR} represent the *transmission tracking*. They account for a change in the phase and magnitude inside the cables, adapters, etc. for the transmitted signal.
- e_{XF} and e_{XR} represent the *cross-talk* or *leakage*. This portion of the signal goes straight from source to load, without reaching the DUT. These terms weigh less on the error model in modern VNAs and can be neglected.

From the analysis of the block diagrams of Fig. 2.5, one can derive the actual S-parameters as functions of the measured S-parameters (s^M_{ij}), the 6 forward error terms (e_F) and the 6 reverse error terms (e_R), that is:

$$s_{ij} = f\left(s^M_{ij}, e_F, e_R\right) \tag{2.21}$$

With the SOLT (*Short-Open-Load-Thru*) calibration, all of the 12 error terms are found in three steps. First, the measured s^M_{30} (s^M_{03} for the reverse model) yields e_{XF} (e_{XR}) by simply connecting port 1 and 2 to the load, since these terms are associated to a port-to-port cross-talk: this reduces the number of model terms to 10. Next, a 1-port calibration is performed as previously shown

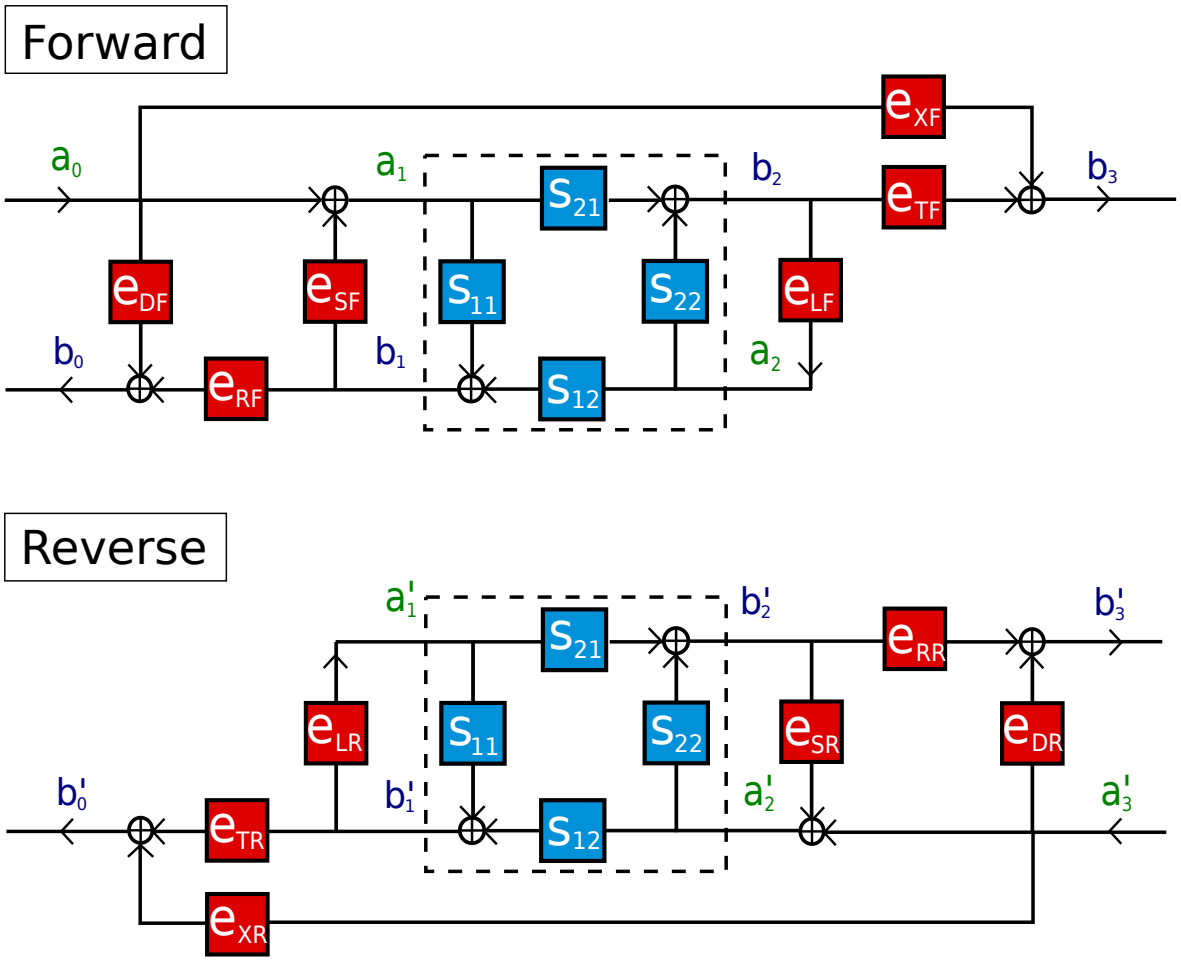


Figure 2.5: 12-term error model.

(SOL calibration), and finally, a direct connection, or *thru*, is inserted between port 1 and 2; s_{00}^{M} (s_{33}^{M}) yields e_{LF} (e_{LR}) and s_{30}^{M} (s_{03}^{M}) yields e_{TF} (e_{TR}). A longer transmission line of length L could also be used as long as its parameters are known in advance. It is therefore evident that for this kind of algorithm (and for the SOL calibration from which it is extended), the knowledge of all the S-parameters of all the calibration standards for all the frequencies of interest is a precondition. This means that the standards have to be fully and unambiguously characterized *before* employing this approach, by means of a comparison to a reference calibration, simulations, etc... Therefore, this approach is intrinsically susceptible to some degree of residual error, due to the lack of full knowledge of the actual behaviour of the standards (caused by real-world standards' non-idealities, such as finite conductivity and non-zero losses, or property modifications at millimeter-wave frequencies). For all these reasons, SOLT is usually employed in the lower part of the terahertz spectrum, in the RF range up to 40 GHz and up to 110 GHz in industry.

8-Term Error Model A 2-port network can be represented in several other manners, including 16-term or 8-term models. The latter derives from the former by neglecting all leakage parameters. Fig. 2.6 describes the associated error model [27]. This model is used when separate receivers are used for all scattered waves and assumes that the switches are perfect (its imperfections can be cancelled out in a four-couplers VNA) and the test-port match does not change when changing their positions. In this model we can spot all the non-idealities introduced thus far:

- e_{00} and e_{33} represent the directivity;
- $e_{10}e_{01}$ and $e_{32}e_{23}$ represent the reflection tracking (reflection loss);
- e_{11} and e_{22} represent the source match;
- e_{22} and e_{11} represent the load match;

- $e_{10}e_{32}$ and $e_{23}e_{01}$ represent the transmission tracking (transmission loss).

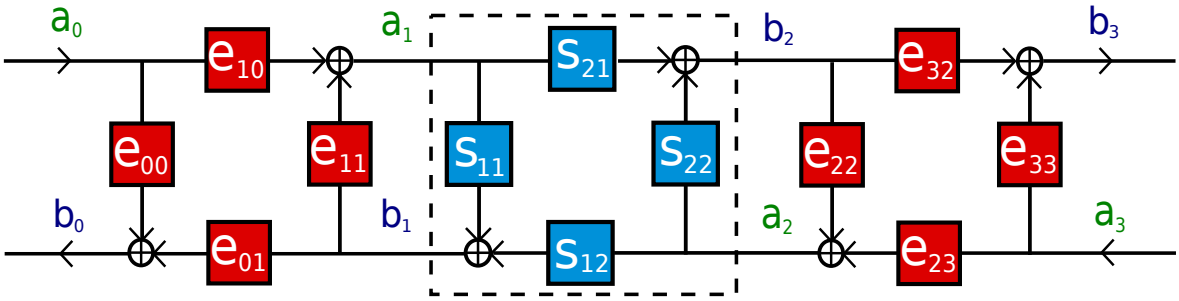


Figure 2.6: 8-term error model.

We can see that all the error terms can be grouped into two sets: error adapter X (the four error terms at the left of the DUT box) and error adapter Y (the four at the right). The cascade configuration makes it trivial to represent the situation using transmission matrices, i.e.:

$$[T_M] = [T_X] \cdot [T_{DUT}] \cdot [T_Y] \tag{2.22}$$

The knowledge of the error terms matrices $[T_X]$ and $[T_Y]$, quantified by a calibration procedure, will allow to retrieve the intrinsic matrix of the DUT, as it can be shown by inverting the previous formula:

$$[T_{DUT}] = [T_X]^{-1} \cdot [T_M] \cdot [T_Y]^{-1} \tag{2.23}$$

Mathematically speaking, for the 8-terms error model only 7 independent unknown are present and it can be demonstrated that just three calibration standards are sufficient and it is not necessary to know the physical description of any standard [95, 29]. A complete description of the TRL algorithm is presented in Appendix B, as it has been implemented by our team. Furthermore, this 8-terms model can even be transformed into a 10-terms model [96] and, by adding cross-talk terms, to a 12-terms: it follows that the terms of each of the presented models are equivalent and interchangeable, as they describe the same system.

Several calibration techniques could fill in all the error terms, and rely on information redundancy to relax requirements on the full knowledge of the standards, and the unknown features are "automatically" computed during the calibration itself: the most common are *Short-Open-Load-Reciprocal* (SOLR) [30], *Line-Reflect-Match* (LRM) [28], *Line-Reflect-Reflect-Match* (LRRM) [22] and *Thru-Reflect-Line* (TRL) [27]. The LRM and LRRM calibrations are sometimes preferred at HF, as their accuracy does not depend on all the standards being fully known, only the time delay of the line and the load's inductance; however, one should also consider that precision resistors are difficult to fabricate in BiCMOS processes [146].

In the TRL calibration, the reflect standard should have a high reflection coefficient (a short or an open are sufficient, but not necessary) and the maximum phase must be known within $\lambda/4$, even if its absolute value can remain unknown. The thru can be of any length, even though a zero-length one will be more precise, since it is lossless and does not generate reflections. The line should present the same characteristic impedance and propagation constant as the thru, but needs to be of a different length. Even so, its characteristic impedance should be well-defined and close to that of the system, i.e. usually 50 Ω. Hence, thanks to these relaxed requirements, the TRL calibration technique allows the use of either non-ideal or not fully known standards. Nevertheless, to cover a large bandwidth, several lengths might be needed: at each frequency, the phase difference between the thru and the line should be greater than 20 degrees and less than 160 degrees. This means that in practice TRL works on a frequency ratio of approximately 8. A simple approximated formula to define an appropriate and flexible frequency range of validity of TRL lines can be the following [44]:

$$f_{\text{min}} = \frac{c}{20\ (\text{L}_{\text{line}} - \text{L}_{\text{thru}})\ \sqrt{\varepsilon_{\text{r,eff}}}}\ ; \qquad f_{\text{max}} = 9\ f_{\text{min}} \tag{2.24}$$

To overcome the limitations on TRL bandwidth, Marks [64] proposed a variation on the TRL calibration called *multiline TRL* (mTRL), which exploits, as the name suggests, multiple lines with different lengths which provide redundant data that are analysed through advanced statistical computations. It is a reliable alternative to SOLT to use for probe-tip calibration, since it appears to be less sensitive to probe misplacement [94] and it is a well-accepted approach at millimeter-wave frequencies, since it limits the uncertainties into the resulting calibration models. However, the evident drawback of die surface consumption, along with requiring additional measurements that may introduce measurement errors (contact issues on aluminium pads, misalignment, etc...), and finally the calibration quality acceptability of our designed single line covering the spectrum up to 500 GHz, made us opt for a simple TRL calibration, instead.

In the following of this manuscript, we will make use of SOLT and TRL calibration, since our designed calibration standards are mainly optimized for that purpose. SOLT is know for being a good broadband solution up to 110 GHz, whilst TRL, which has a limited frequency range and cannot go down to low frequencies due to the employment of finite-length lines, has, on the other hand, very few systematic errors [132] and can be used at HF where it proves more accurate, since it does not rely on the lumped nature of the equivalent circuit of the standards, which at frequencies beyond 110 GHz may not be valid any more [130, 66]. Moreover, TRL calibration performs "intrinsically" better, since it allows to set the reference plane in a transmission line where wave parameters, voltages and currents can be rigorously defined [130]. Finally, since transmission lines with acceptable losses are fabricated in a nanoscale BEOL with multiple copper layers, both TRL and mTRL calibrations, which have been already largely employed in III-V semiconductor device characterization, are being routinely applied to silicon-based technologies as well [146].

2.2 On-Wafer Measurements at High Frequency

Historically, the characterization of millimeter-wave (mmW) devices is made in a rectangular waveguide (WG)-based system, in which the DUT package is connected through a WG to the VNA. The resulting measurements are free from ambiguity and spurious contributions since only a single mode propagates from the interface of the WG. However, this technique turned out to be time consuming and costs were high. Also, presenting modern VNAs usually coaxial test-ports in their front side (above 50 GHz, both coaxial and waveguide are possible interfaces, while only rectangular WG dominates above 110 GHz), often times high frequency extenders (discussed in the following) realized with WG interfaces are employed.

That said, nano-electronic devices and integrated circuits are, as for them, realized on a planar environment, being that, when dealing with model extraction and validation, it is preferred to maintain the DUTs in their original wafer substrate. For this reason, on-wafer RF probes have been adopted in order to measure HF devices without the need of prior packaging [71] and transforming the measurement interface from waveguide or coaxial to planar [41].

2.2.1 Radio-Frequency Probes

RF probes consist of a body with the instrumentation interface, sometimes followed by a waveguide flange and/or a transition to a micro-coaxial cable, with a final transition to a planar waveguide (e.g. CPW or microstrip), and a possible probe-to-DUT interface: the probe tips, made of different materials (tungsten, beryllium-copper, gold-nickel alloy...). Some WG-based probes also integrate a *bias tee*, which is used in active device measurement to provide a DC bias to the transistor: it presents a low pass frequency response by providing a direct low resistance DC path for supplying DC current and voltage to the DUT [21]. Therefore, probes work as a transition from the

three-dimensional medium (coaxial cable or rectangular waveguide) to two-dimensional (coplanar) probe contacts, and an electro-mechanical contact is made at the DUT level through the pads [92].

The pads are three conductive contact platforms typically arranged in a ground-signal-ground (GSG) configuration (a symmetrical topology with the purpose of reducing any electromagnetic fringing line from the signal pad to the substrate), and made of soft metals (gold or aluminium) to allow low-ohmic connection without damaging the probe tips [41].

Compared to direct DUT package connection to the measuring systems, RF probes are more prone to user-led errors such as poor pad contact and placement, bad planarization, etc... and particularly to HF systematic errors. The waves travelling in a coaxial (or WG) section have to be properly converted into a planar field distribution when reaching the pads, by taking care of the different propagation modes. As we will see in the next chapter, the probe tips act as a discontinuity to the signal path and generate higher propagation modes. Therefore, the on-wafer lines should support only a single quasi-TEM mode and exclude higher-order ones [92].

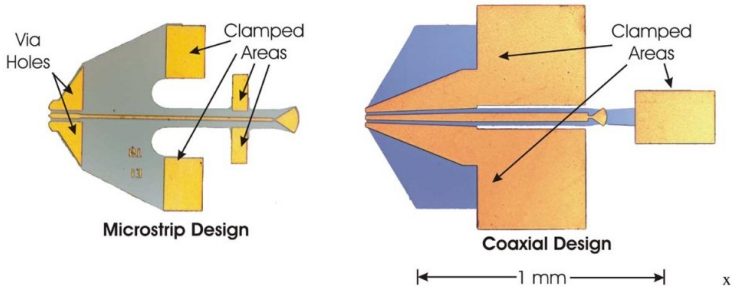


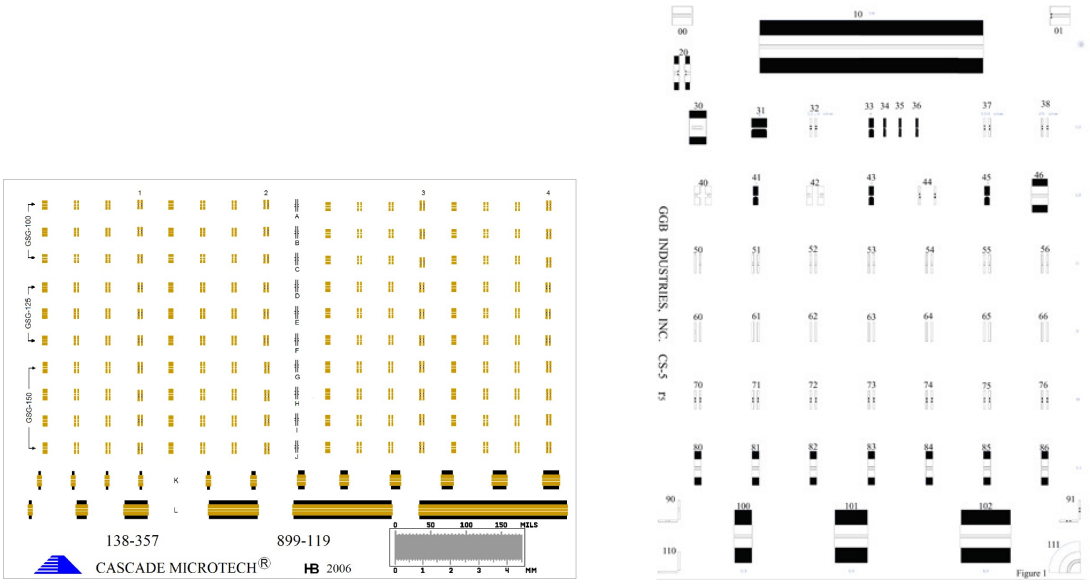
Figure 2.7: Photomicrographs of a micromachined on-wafer probe with two transmission line designs. Courtesy of [88].

In recent years, emerging THz technologies have gone hand-in-hand with new probing systems, which now extend beyond 1 THz [6] and allow ever-reducing pitches (hence, pad sizes). New probe design and fabrication include micromachined and silicon-microfabricated technologies [87, 89, 88, 67] (Fig. 2.7) which reduce contact resistance and other parasitics and may embed elaborate RF circuitry into the on-wafer probes, allowing novel metrological techniques such as HF differential probing (integrated baluns) [149, 150], on-wafer temperature (Schottky diodes) [139] and strain sensing [148]. Other solutions consist in active probes incorporating an IC based on non-linear transmission lines (working as a VNA) and CPW tips [91, 137].

2.2.2 Calibration on a Dedicated Substrate

As we have seen, all the contributions from the measuring setup up to the planar environment of the DUT (cables, connectors, extenders...) are part of what we called "extension" of the VNA test ports and the associated systematic errors are removed together with the ones produced by the VNA by calibration. It is a common practice to extract the figures of merit of a device at relatively low frequency (typically 20 GHz) by some dedicated piece of equipment to perform (a two-tier) calibration (usually SOLT, but also LRRM) [73, 132]. In such structures, the calibration standards are located in a planar support called *calibration substrate*. The standard manufacturing has of course fabrication tolerances, for this reason the standard response will deviate from its predicted one determining a *residual calibration error* (also called *characteristic data*), which is described in the data sheets and may be provided inside the network analyser's software for a complete correction. Some examples of calibration substrates are shown in Fig. 2.8.

The calibration and de-embedding techniques are sometimes (e.g. at HF) less critical than accurately minimizing the parasitic resistances, capacitances and inductances of the calibration and de-embedding standards, since the measurement and de-embedding error is proportional to



(a) 138-357 calibration substrate for the 220-325 GHz range (courtesy of FormFactor Inc. – Cascade Microtech)

(b) CS-5 calibration substrate for up to 220 GHz (courtesy of GGB Industries Inc.)

Figure 2.8: Two adopted calibration substrates.

those parasitics [146]. In theory, in fact, when a system is calibrated, the calibration is specifically only valid if nothing is changed in the setup except the DUT: otherwise, serious over/under estimations of the performance of the structures/devices may take place early with frequency. It becomes more relevant at high frequency, when a small change of the error coefficients computed by calibration can have a strong impact on the measurements of the DUT S-parameters.

When we employ calibration substrates, we are prone to those changes of the error coefficients. In this case, in fact, the standards often come embedded into a ceramic substrate (e.g. alumina or fused silica) instead of silicon-based materials (Si, SiO_2, GaAs...), in form of (typically) gold patterns, which do not provide the same contact resistance of the aluminum pads of the DUT. Moreover, when laying on a different substrate during measurement, a different probe-to-substrate interaction will occur in the host medium, providing a residual error (*coupling*) physically contributing as an additional capacitance which is a function of the permittivity, and gets higher at HF [38]. In Fig. 2.9 we can observe the interaction of the field generated by the probes with two different environments; electro-magnetic simulations with similar setups have been carried out in the past by our team [33], with which we will not deal in detail in this work. In an attempt to improve the off-wafer measurements, post-calibration optimization by simulation can be performed to correct the calibration deviation induced by mmW phenomena: in this way, however, the calibration process becomes evidently cumbersome [124].

Therefore, few research centers like ours embed calibration standards in the same silicon environment with the same material and shape of the pads for improved accuracy of measurement. The resulting calibration is in all respects, *"on-wafer"*, which means on the same silicon substrate of the measured DUT. For this reason, on-wafer calibration kits have been developed and considered in the following of this work.

However, one major issue related to the employment of on-wafer calibration is the difficult contact repeatability on aluminium contacts [132, 130] which oxide quickly and good connection depend on the probes employed, the number of previous contacts and the skill of the operator. Williams *et al.* tried to solve this by proposing a gold-plating process to improve the measurements comparable to the quality of those made on commercial ISS' gold pads [130]. Moreover, although the loss in the SiGe BiCMOS' BEOL is considerably lower than, for instance, a CMOS' digital BEOL, it is still a factor of three larger than that of the coplanar transmission line on an ISS, and this may have serious implications in some of the most high-sensitive figures of merit of an HBT, e.g. the

f_{MAX} and the minimum noise figure [146].

It is also worth mentioning the use of electronic calibration techniques for microwave characterization [119] that have been allowing for two decades now to eliminate the requirement of multiple standards placement and minimize operator interactions, thus reducing measurement uncertainty and calibration process time. They employ solid-state circuit technologies and solutions have been proposed very recently for sub-millimeter frequencies with Schottky diodes [140]; feeding the diodes with a varying bias voltage, it is possible to generate a set of different impedances, thus reflection coefficients, to retrieve the values of the error terms of the system. The electronic calibration relies on an accurate definition of the DC bias point, yet it has a limited coverage over frequency, it is more costly to implement and yet not widely-established.

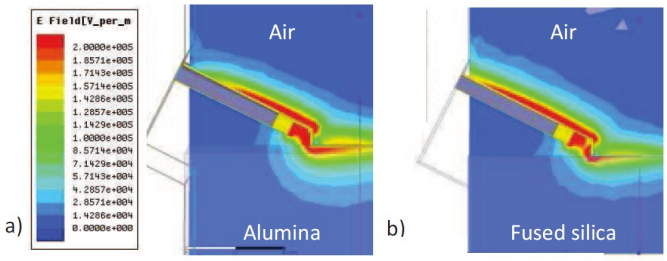


Figure 2.9: Side view of the simulated electric field (HFSS) describing the interaction between GSG probes and a transmission line built on different substrates, at 110 GHz. From [38].

2.2.3 De-Embedding Routines

The performed calibration is needed to establish the reference plane, a fictitious three-dimensional *locus* to which the measurements relate, at the edge of the extended test ports. In the case of planar standards, this location is not straightforwardly identified in many cases, such as for the SOLT calibration: the plane is ambiguously located somewhere at the probe tips, i.e. at the interface of the extended test ports [132]; indeed, it can also be demonstrated that this calibration technique is highly sensitive to the probe position [99].

For TRL, on the other hand, the reference plane is set at the center of the thru and translated back through a mathematical transformation to a precise location along the access lines, only by the knowledge of the propagation constant and the length of the line [135]. In this way, the plane can be set sufficiently far from the problematic probe-to-pad discontinuity, where the assumption of just a single mode propagating cannot hold true. Eventually, thanks to the TRL calibration, all the non-idealities generated by the EM field scattering at the interface are excluded.

Some notorious parasitic effects and multi-mode propagation due to the RF probes are *substrate coupling* (a capacitance; the value of this is largely reduced by on-wafer calibration [38]), *probe-tip radiation* and *probe-to-probe coupling*, and they add up to the systematic errors due to the "mechanical" handling (bad contact, etc...). However, also other types of calibrations (such as SOLT) are perfectly acceptable in practice and provide satisfying results, particularly at lower frequency, but it is true only when the probe tips are electrically small, namely their dimension is much smaller than the wavelength of the generated signals.

However, while these effects are related to the RF probes and can be largely excluded by calibration, other parasitic effects take place after the reference plane. In fact, from the probe tips many extra metal connectors and vias are needed to provide a signal path to the device buried in silicon. This entire structure is dubbed *back-end of line* (BEOL), as opposed to the *front-end of line*, i.e. an active device, for example. Parasitic contributions rise between and along these connections in the form of resistances, capacitances, inductances... and since the reference plane is set at the pad level, and the pads are usually around 20 times larger than the device accesses, all the spurious effects are comprised in the calibrated measurement. The demand of high accuracy on-wafer parameter extraction becomes even more stringent at HF, as we will discuss further on.

Eventually, one may want to apply a second-tier calibration to remove (or *de-embed*) those effects from the first (actual) calibration: hence, we call this procedure *"de-embedding"*. De-embedding consists in pushing the reference planes closer to the actual device down to the terminals at the bottom metal layer, by removing, for example, the effects of test fixture and metal layers composing the BEOL. Highly accurate de-embedding schemes become a necessity, but complexity does not necessarily mean precision, since the more complete the procedure aiming to get better results is, the more additional errors will be introduced due to additional measurements and the more time will be spent performing it. The accuracy of the method depends on the DUT and its surroundings, and a technique proving fine for a device might not give correct results for another.

Throughout time, over 450 methods, based on both electrical models and mathematical expressions, have been published to remove the effects of components that obscure the device response and behaviour [97] and still it does not exist an established de-embedding method for silicon devices at frequencies above 110 GHz. The first, most basic de-embedding method which was established [126] was a simple "open" de-embedding scheme (or *standard*, or *dummy*). The goal of this technique was to evaluate the pad capacitance through an open test structure located at pad level. If one also wants to include the interconnect resistance and inductance, another standard can be added, i.e. a short, to finally perform a *"open-short"* (or *"short-open"*) de-embedding down to DUT level [55] by keeping the same BEOL configuration as the DUT one wishes to de-embed (i.e. the transistor). This method which relies on a lumped model of the access (constant capacitance and inductance over frequency) is largely sufficient at low frequency [1, 25] and is considered as the industry standard. However, there exist many of this approach alternatives, some of them trying to capture the distributed nature of parasitic at higher frequencies and are more sophisticated. They are either based on elaborate lumped circuits, or transmission line theory, or a four-port error model. Eventually, however, simpler alternatives are generally preferred.

According to the order of precedence in using the de-embedding stardards (short and open), two distinct models can be drawn. The circuit models which take into account the parasitic elements are shown in Fig. 2.10. In this two-step evaluation of parasitics, the short standard will be used to evaluate the impedances Z_{P1} and Z_{P2}, at port 1 and 2 respectively, as well as the mutual impedance Z_M while the open standard will be for the admittances Y_{P1} and Y_{P2}, at port 1 and 2 respectively, plus the coupling admittance Y_C.

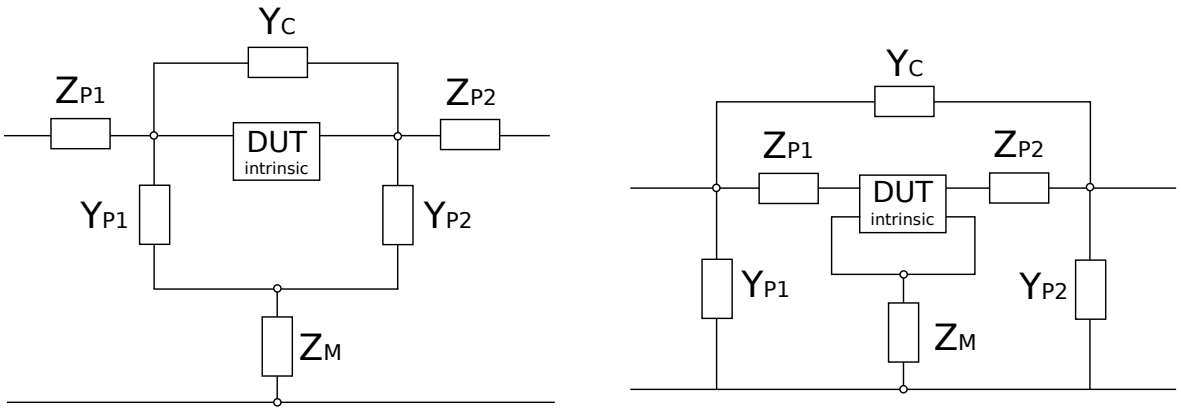


Figure 2.10: Equivalent two-step circuit model for short-open (left) and open-short (right) de-embedding.

Rigorously, each de-embedding step will provide a matrix representation of the parasitic contribution. Short structures yield a **Z** matrix, open structures yield a **Y** matrix, while lines (if any, but not present in this simple representation) yield ABCD chain matrices. The serial parasitics' matrix based on the equivalent T-model, and the parallel parasitics' matrix based on the equivalent Π-model are defined by, respectively:

$$\mathbf{Y}_O = \begin{bmatrix} Y_{P1}+Y_C & -Y_C \\ -Y_C & Y_{P2}+Y_C \end{bmatrix}; \qquad \mathbf{Z}_S = \begin{bmatrix} Z_{P1}+Z_M & Z_M \\ Z_M & Z_{P2}+Z_M \end{bmatrix} \tag{2.25}$$

As a way of example, the short-open de-embedding algorithm (Fig. 2.10, left) can be developed as follows. First, the S-parameter matrices need to be converted to Z-parameters, in order to linearly remove the contribution of the impedances, namely:

$$\begin{aligned} \mathbf{S}_{\text{dev}} &\rightarrow \mathbf{Z}_{\text{dev}} \\ \mathbf{S}_{\text{O}} &\rightarrow \mathbf{Z}_{\text{O}} \\ \mathbf{S}_{\text{S}} &\rightarrow \mathbf{Z}_{\text{S}} \end{aligned} \tag{2.26}$$

in which the "$_{\text{dev}}$", "$_{\text{O}}$" and "$_{\text{S}}$" subscripts indicate the calibrated DUT, open de-embedding standard and short de-embedding standard, respectively. The next step is to mathematically remove the impedances to both the DUT and the open standard:

$$\begin{aligned} \mathbf{Z}_{\text{dev,S}} &= \mathbf{Z}_{\text{dev}} - \mathbf{Z}_{\text{S}} \\ \mathbf{Z}_{\text{O,S}} &= \mathbf{Z}_{\text{O}} - \mathbf{Z}_{\text{S}} \end{aligned} \tag{2.27}$$

where $\mathbf{Z}_{\text{dev,S}}$ is the device measurement de-embedded from the short and $\mathbf{Z}_{\text{O,S}}$ is the open measurement de-embedded from the short. These resulting matrices are now converted into Y matrices much like before and eventually, the Y matrix resulting from the difference of the two is transformed back to S-parameters, i.e.:

$$\mathbf{Y}_{\text{dev,SO}} = \mathbf{Y}_{\text{dev,S}} - \mathbf{Y}_{\text{O,S}} \rightarrow \mathbf{S}_{\text{dev,SO}} \tag{2.28}$$

where $\mathbf{S}_{\text{dev,SO}}$ is the device measurement de-embedded from the short and the open. The dual algorithm can be applied for a open-short de-embedding.

2.2.4 Characterization at Millimeter Wave Frequencies

Millimeter-wave (mmW) and terahertz (THz) measurements are challenging and few authors engaged in the characterization of transistors on silicon substrates at those extremely high frequencies. Voinigescu *et al.* and Deng *et al.* [121, 24] performed analysis and validated the compact model up to 330 GHz on the same silicon technology presented here, but also on mmW circuits (amplifier and VCO) [121] with different off-wafer calibration techniques. Williams *et al.* performed multiple on-wafer calibrations (and de-embedding techniques) but stopped at 110 GHz. And even though Galatro *et al.* and Fregonese *et al.* [40, 32, 33] reached 325 and 500 GHz, respectively, they did not benchmarked any measurement comparison with actual probe tip models, like it will be shown in the following of this work.

These measurements represent a difficult task from both the device and the hardware point of view. As a matter of fact, the majority of VNAs works in a limited range of frequencies (usually up to 67 GHz), which is sufficient to cover the majority of HF measurements of devices and systems. Monolithic microwave integrated circuits (MMICs) and mmW devices, however, are designed for applications above 110 GHz, where many network analysers alone cannot perform measurements. Also, coaxial cables involved in carrying the signals to measure the HF parameters need to be adapted to these elevated frequencies. It is the dimension of the outer and inner conductors that allow the transmission of a single transverse electromagnetic mode (TEM) propagation, and to cope with higher frequencies, the dimensions have to be reduced. Using coaxial cables alone to direct the signal from the VNA can be technically done only up to 145 GHz for broadband applications (0.8-mm coaxial connector [4]). Higher signal losses and distortions are inevitable inside the coaxial cables at these frequencies; this translates to a degradation of measurement and consequently of the calibration procedure. At higher frequencies, direct connection to RF probes can be performed up to 220 GHz with the use of a waveguide output [4].

Above the limits of direct VNA generation of broadband signals, however, it is necessary to extend the frequency of the VNA in selective ranges through *frequency extenders*, also known as

mmW heads. These are transmitter-receiver modules embedded with a reflectometer and circuitry for multiplication and sub-harmonic mixing and presenting a waveguide or coaxial output, depending on the extended range. Waveguides are the ones that limit the application of extenders to specific ranges, since their dimensions should allow the field to propagate as a single transverse electric (TE) mode (TE_{10}), between the cutoff frequency and the appearance of a second mode. The extension modules are available up to 1.5 THz nowadays [120]. The conversion to a quasi-TEM mode, which is the one propagating on a planar line (microstrip or CPW), is then performed by a short section of micro-coaxial cable or directly through a microstrip membrane or a micromachined silicon CPW [92].

Another issue related to sub-millimeter and millimeter wave frequency measurement is power control. Performing a characterization on an HBT requires the user to combine a DC biasing and a low RF signal (through a "bias tee"). The user has to determine which RF level has to be input by first visualizing the DC characteristics (Gummel plot) without any RF interference (system on "hold" mode) and subsequently superimpose a RF signal with adequate amplitude ("continuous" mode), at risk of major distortions if it is too high. A too large RF source power consequently distorts the performance of the HBT with an incorrect bias. In our measurements, we take the typical output power of the mmW heads as reported by the manufacturer and use a mechanical attenuator to reach power level ranging from -30 to -35 dBm.

However, low RF power may result in poor stability performance and power drift, for two main reasons [109]: 1) source power is frequency-dependent and tend to drop down to too low levels or vice-versa, since the ALC inside the VNA (in charge of controlling the input power level) is bypassed when extenders need to be used, too high values (which is a more critical scenario, since this affects the DC characteristic). This results into uncontrolled and fluctuating power at probe tips (as we have often experienced in our 110 GHz measurement setup); 2) the received power is not calibrated, making it prone to (even huge) losses due to the DUT characteristics (inconstant dynamic range), leading to very low SNR, thus significant measurement uncertainties. These aspects may have a non-negligible effect when driving an active device, as the user may experience oscillations and dips in the measured S-parameters, although we took special care into verifying the stability of our input and output power level. Solutions to monitor and control the source power may be adopted [109].

Apart from the measurement setup, HF measurements are critical from the device point of view. Parasitic effects of the DUT are superimposed to the device characteristics, appearing at mmW frequencies and potentially worsening the device performance, hence the whole electronic circuit performance, too. Each object or device can be modelled as a lumped model as long as it is electrically small, that is, its spatial extend is smaller than approximately one-tenth of a wavelength [52]. At HF, the signal wavelengths become comparable or smaller than the devices where they are injected: at mmW frequencies, only distributed circuit element models accurately describe the device and highly reflective standards, like the dummies used for de-embedding, are difficult to probe. That is a major challenge, since de-embedding accuracy relies on the probing quality of the de-embedding structures. The de-embedding assumption of a lumped-element circuit approximation for a simple two-step approach may therefore weaken the validity of the measured results at high frequency [117]. Also, the complex layer architecture of mmW devices, presenting a sequence of metal and dielectric stacks, degrades the accuracy of the extracted model parameters. Finally, propagation of the field inside the substrate is another common issue [43].

2.2.5 The Adopted On-Wafer Measurement Setup

All the results showed in this work come from measurements performed on four different bands, with different setups, from 1 GHz up to a maximum frequency of 500 GHz, and are often correlated to simulation as it will be described in the following chapter. An Agilent (now Keysight Technologies) N5250A module [112] (see Fig. 2.11a) is used for the first band and it is composed by an E8361A PNA Network Analyzer, covering internally the range from 10 MHz to 67 GHz [113], a combiner, a bias tee, and the N5260A millimeter head controller [114] for two test heads providing a

	E8361A PNA	ZVA24
Frequency Range (GHz)	0.01-67	0.01-24
Test Port Connector	1.85 mm	3.5 mm
Dynamic Range (dB)	>94 (0.75-67 GHz)	>125 (0.7-24 GHz)
Directivity (dB)	>34 (2-67 GHz)	>40 (0.7-24 GHz)
Source Match (dB)	>34 (2-67 GHz)	>36 (0.7-24 GHz)
Reflection Tracking (dB)	<0.09 (2-67 GHz)	<0.1 (0.7-24 GHz)
Load Match (dB)	>34 (2-67 GHz)	>40 (0.7-24 GHz)
Transmission Tracking (dB)	<0.15 (2-67 GHz)	<0.1 (0.7-24 GHz)
Output (typ.) (dBm)	-27 to -7	-40 to +16

Table 2.1: VNA specifications as provided by [113] and [104]. System data are given after system error correction (calibration).

	N5250A (module)	ZC220	ZC330	ZC500
Frequency Range (GHz)	0.01-110	140-220	220-330	325-500
Test Port Connector	1 mm	WR5	WR3	WR2.2
Dynamic Range (typ.) (dBm)	68 to 120	115	115	105
Directivity (typ.) (dB)	N/A	>25	>20	>20
Source Match (typ.) (dB)	N/A	>25	>20	>20
Output (typ.) (dBm)	-22 to -2	+1	-7	-11

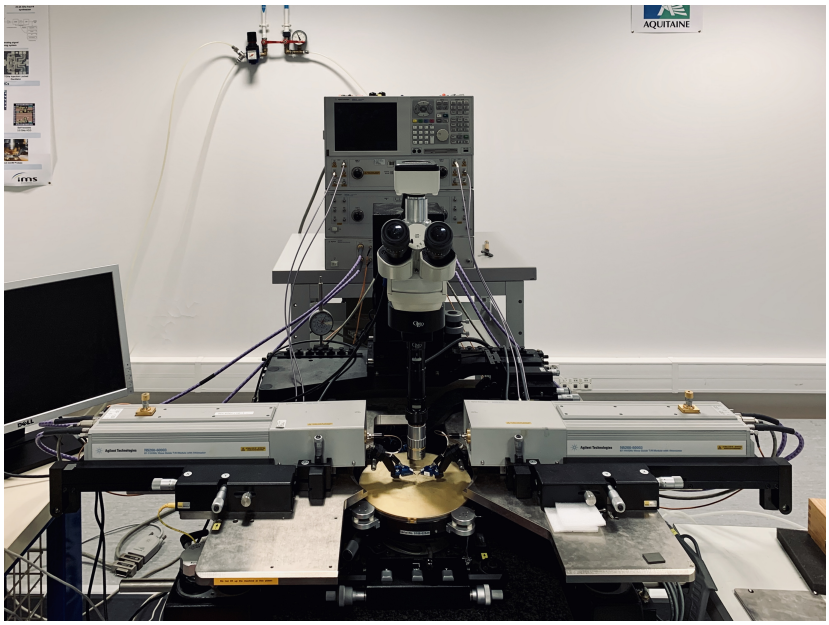
Table 2.2: Frequency extenders specifications as provided by [112] and [106]. System data are given before system error correction (calibration).

signal from 67 to 110 GHz. For the three remaining bands, another VNA is exploited. It is the ZVA24 from Rohde & Schwarz, with nominal operation up to 24 GHz [104]. To span in the WR5 (140-220 GHz), WR3 (220-330 GHz) and WR2.2 (330-500 GHz) waveguide bands, three different frequency extenders again from Rohde & Schwarz are adopted, respectively the ZC220, ZC330, and ZC500 [106] (see Fig. 2.11b). Validity frequencies, dynamic ranges and the system data as listed by the vendors are compared in Table 2.1 and Table 2.2 for the VNAs and the extenders, respectively.

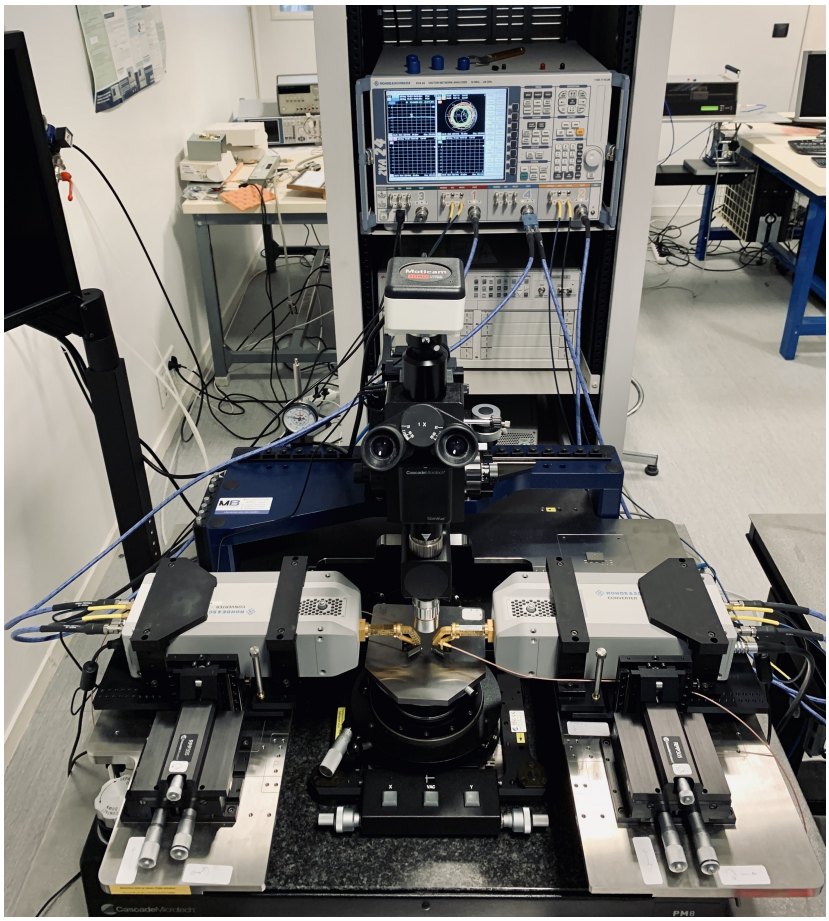
For each band two dedicated probes are used to contact the DUT, coming from different vendors (shown in Fig. 2.12 and 2.13). In the band up to 110 GHz, we dispose of two GGB Industries Inc. Picoprobes® with pitches of 100 µm and 50 µm: for a better readability in the text, they will be referred to as "PP-110" and, since we have not measured with the 50 µm pitch one in this band, we will always refer to the 100 µm in the following [48, 10]. In the 140-220 GHz band, a Picoprobe with 50 µm is used along with a Cascade Microtech Inc. Infinity Probe® with a pitch of 100 µm, and they will be referred to as "PP-220" [49, 37] and "IP-220" [31, 13], respectively. In the 220-330 GHz band, a Picoprobe with 50 µm is again employed along with another Infinity Probe with a pitch of 50 µm, and they will be referred to as "PP-330" [50, 36] and "IP-330" [31, 12], respectively. Finally, for the 325-500 GHz band, a Picoprobe with 50 µm has been once again adapted, plus a Dominion MicroProbe Inc. T-Wave Probe® with a pitch of 50 µm, and they will be referred to as "PP-500" [51, 35] and "TW-500" [31, 89, 88], respectively. TW-500, based on a micromachined silicon technique has not been exploited for this thesis, but the probe design will be presented as a modern advanced manufacturing example.

Table 2.3 sums up all the main characteristics of the used probes as declared by their respective vendors.

Picoprobe probes have a coaxial 1 mm connector just for the lowest band (that matches our N5250A module output), while the others have waveguide inputs due to the higher frequencies involved. Then, inside the probe body, either a low-loss cable (Fig. 2.14a) connects the coaxial input to the microcoaxial line, or the signal from the waveguide is collected by a plunger (Fig. 2.14b) emerging from the input of the microcoaxial cable. At the opposite end, the microcoaxial is shaped out to form the signal tip with ground tips soldered to the outer connector. The DC bias tee is em-



(a) Probe station for 110 GHz measurements. The N5250A module is visible: E8361A PNA, mm head controller, 67-110 GHz frequency extenders, DC supply, mounted PP-110



(b) Probe station for 220-500 GHz measurements. The ZVA24 and DC supply are visible, as well as ZC330 frequency extenders and PP-330

Figure 2.11: RF probe stations.

bedded in the probe body when the probe input is not coaxial.

Infinity probes, on the other hand, introduce a thin-film technology at the tip (Fig. 2.14c). After

Figure 2.12: Pictures of the used probes for millimeter wave measurements. From top left to bottom right: IP-220, PP-220, PP-500, TW-500.

	PP-110	PP-220	IP-220	PP-330	IP-330	PP-500	TW-500
Frequency Range (GHz)	1-110	140-220	140-220	220-325	220-330	325-500	325-500
Pitch (µm)	50/100	50	100	50	50	50	50
Probe Input	coax 1 mm	WG WR-5	WG WR-5	WG WR-3	WG WR-3	WG WR-2.2	WG WR-2.2
Tip Realization	microcoax	microcoax	microcoax-microstrip	microcoax	microcoax-microstrip	microcoax	tr. line on micromachined Si
Tip Material	BeCu	BeCu	Ni alloy	BeCu	Ni alloy	BeCu	Ni
Tip Length (µm)	400	200/250	N/A	250	N/A	150	350
Min Wavelength (mm)	2	1.4	1.4	0.9	0.9	0.6	0.6
Insertion Loss (typ.) (dB)	1.5	2	5.2	3	6.5	4	4.5
Return Loss (typ.) (dB)	15	15	13	15	13	15	15

Table 2.3: Main features of the adopted probe sets.

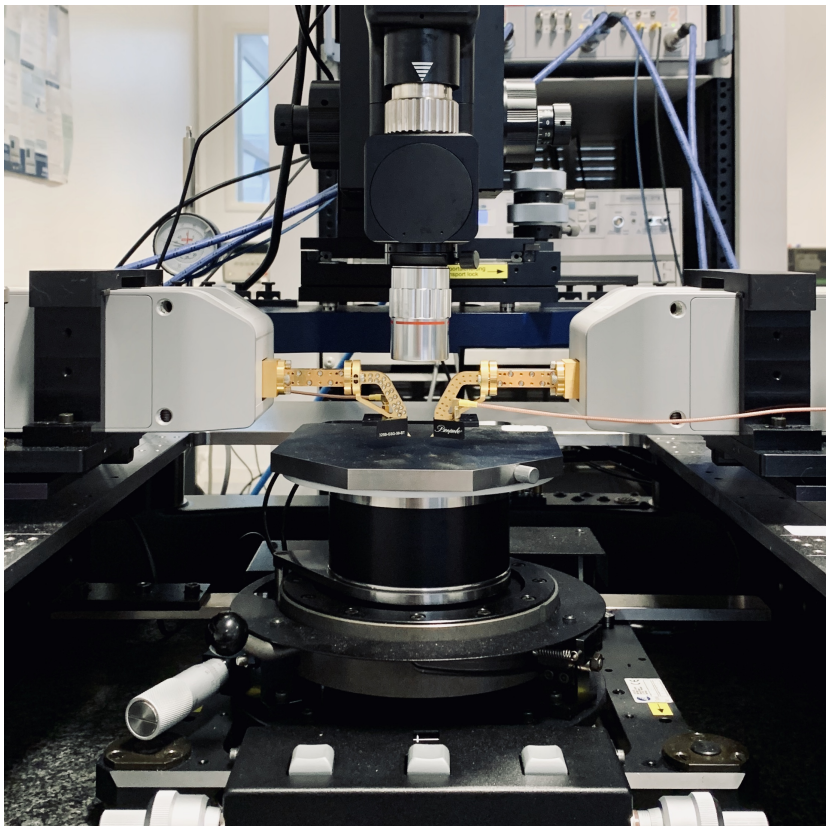
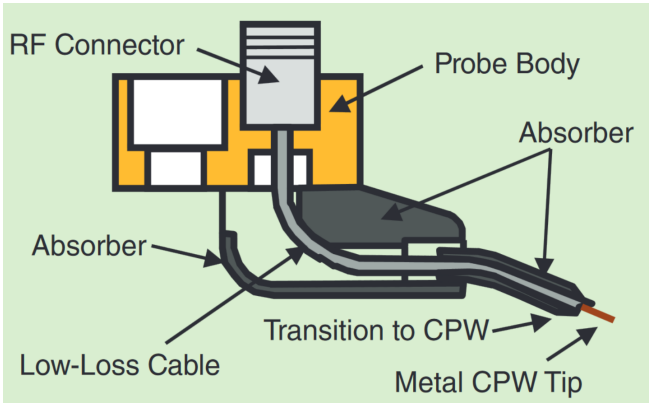


Figure 2.13: Detail on PP330 mounted on the 220-500 GHz probe station.

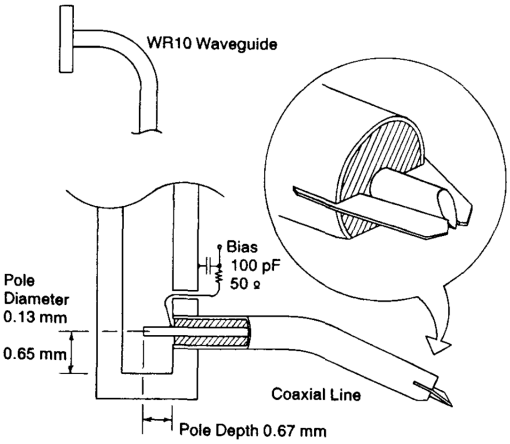
the usual WG-to-microcoax, a polyamide microstrip line is directly attached to the microcoaxial cable, which makes contact with the DUT through CPW tips: introducing an intermediate microstrip is claimed to better confine the signal energy and reduce coupling with the substrate, since the EM field shape of the microstrip is similar to the one of the CPW.

Finally, for the upper band, we adopted a recent probe design: a micromachined probe [87] (Fig. 2.7). It consists of a substrate-supported transmission line (rectangular coaxial line or microstrip) mounted on a silicon substrate, and housed in a clamped metal WG channel. The transition from the rectangular WG is made through a radial stub, connected to a microstrip/coaxial transmission line laid on a 15 µm silicon substrate with gold conductors. The channel eventually transitions to the unenclosed CPW tips.

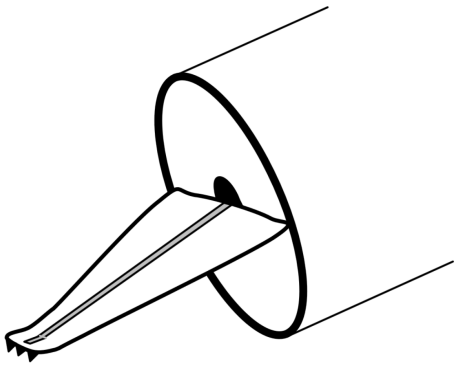
The dedicated portions of wafer, in which the dies with the embedded test devices and on-wafer calibration standards are located, are fixed to the probe station chuck by applying vacuum in order to prevent shifting or flipping as a result of the probe mechanical contact.



(a) Low-loss microcoaxial cable (from [92])



(b) Waveguide to microcoaxial (plunger) (from [60])



(c) MS line membrane probe tips (from [98])

Figure 2.14: Three different RF probe technologies: transitions from connector to wafer.

Chapter 3

Evaluation and Optimization of Layout Design

Contents

ONCE INTRODUCED the general concepts of a measurement system with particular focus on our bench stations and calibration techniques, we present in this chapter the devices and structures under test, whose layouts and BEOL have been designed by our team in two subsequent versions.

The first one embeds elements from previous layouts, such as slots in the metal ground, an aligned configuration, and a specific BEOL, while the second version tries to make up for the problems encountered after analyzing the measurements, featuring a brand new design that will be described in detail. Each structure under test will be presented as it is in the first production run, and any new features added in the second will be described and schematically listed. Later on, measurements and simulations from the two runs will be juxtaposed and throughoutly commented.

The TRL calibration algorithm used by our team, i.e. our calibration "toolkit", will be introduced and particular attention will be given in evaluating the validity of the approximations of the TRL calibration's impedance correction method that we have implemented.

Furthermore, the two calibration methods that were introduced in a general way in the previous chapter, TRL and SOLT, will be compared on actual DUT measurements and simulations. We will see how SOLT calibration deviates when the measurement substrate is different from the silicon where our DUTs are built.

For this purpose we will use the EM simulations carried out on HFSS and an innovative method of characterization of the transistor performances that takes into account the effect of the removal of the probes and the measurement setup: the EM + HICUM co-simulation.

Finally, a complete and original study will evaluate, through both measurement and simulations, with the most complete configuration possible (i.e. the one with all the adjacent structures), to what extent the layout variations of the second run have an impact on the measurements.

3.1 Masks Presentation

In order to make the on-wafer calibration possible, two fabrication masks have been fabricated on several wafers by STMicroelectronics once our team had designed the test structures. Each wafer is embedded with multiple dies, each of them presenting a portion specifically dedicated to calibration test structures and the transistors, which have been intensively measured and characterized to compile this work. These two versions (or *production runs*) have been temporally fabricated one after another, the first in early 2017 and the second in late 2018, as a way of improving the first run after more than one year of characterization and verification of the calibration performance. To respect the order of their creation, they will be named "run 1" and "run 2", respectively, in the following.

On these runs, transistors share the same fabrication process, i.e. STMicroelectronics' BiCMOS 55-nm (B55) technology [16]. All the structures lay on the common silicon die substrate and each technology consists of 8 copper layers (sequentially shortened M1, M2, ...M8) plus one aluminium cap layer above M8, where the probes can be placed, and corresponding copper vias (also shortened V1, V2, ...V7) in the back-end of line (BEOL), that allow access to the transistor (Fig. 3.1). On top of M8, which is used for the microstrip line design, aluminum ground-signal-ground (GSG) contact pads are realized. Metal layers are surrounded by multiple silicon dioxide (SiO_2) layers with different dielectric constants ε_R (approximately 30 passivation layers among which the 8 metal layers are located, and a silicon nitride layer in the topmost part of the stack, extending above the pads) whose harmonic average yields a total $\varepsilon_R \approx 4$.

The evolution of interconnection processes tends towards an ever greater reduction of metal thickness, related to the increasing integration of active devices. For frequencies above few GHz, each µm of the BEOL has a significant influence on the electrical behavior of the devices. For this reason, we will draw particular attention to this aspect, and present distinctions between the two considered runs. The differences, however, come not only from the different metal interconnects, but also from the general layout and topology of structures we designed, as we will describe in this chapter.

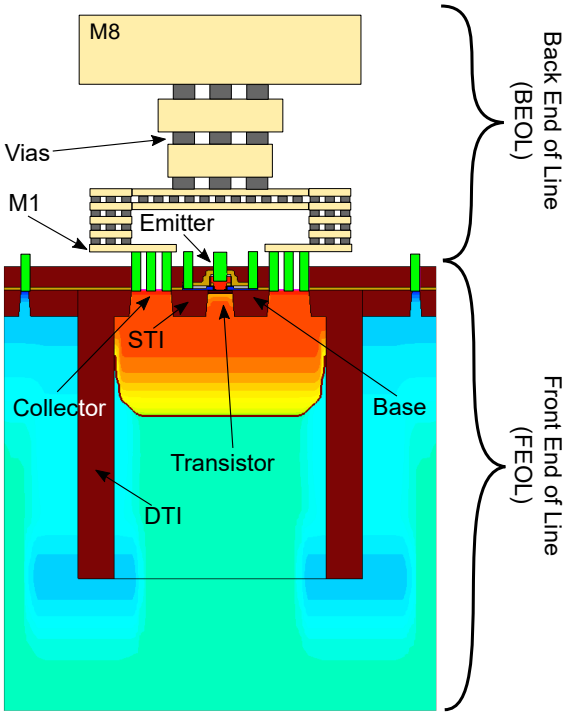


Figure 3.1: Artwork of the sectional view of SiGe BiCMOS technology. Only the BEOL connected to the collector is shown for clarity. Not in scale.

3.1.1 Production Run 1

As presented by Deng *et al.* [23], the test structures for the on-wafer TRL calibration and HBTs in the first production run considered here, have been designed drawing inspiration from a previous design that was implemented on Infineon's B11HFC technology [8]. Fig. 3.2 depicts a floorplan view of the test structure of run 1. Letters and numbers can be used to identify each of the standards on the die at the crossing of an imaginary line drawn vertically and horizontally. Table 3.1 can be used to name the mapped structures.

Run 1 presents a 10-µm-thick silicon dioxide (SiO_2) slot or "ring" extending from the silicon substrate to the top of M8 and surrounding all the test structures for an electrical isolation [23] (shown in aqua color in Fig. 3.2). We reused these elements inherited from previous designs from ST –in which they used it as isolation material– due to lack of clear evidence of their impact. The metal blocks among structures (which make the dielectric in the external proximity of each structure a "ring" indeed) were added to maintain inter-structure distance and additionally comply with design rules in terms of metal density. In [131], it is indicated as a "rule of thumb" to avoid the use of such slots in microstrip grounds for calibration structures, since they may generate excess coupling between structures. However, this recommendation has not been studied in detail by simulation with probe models, as it will be done in the following. Moreover, oftentimes this advice does not seem to be adopted in the industrial environment.

Signal and ground RF pads, mounted on M8, allow hosting both 50 and 100 µm-pitch probes, a necessary condition for measurement up to 500 GHz with our probe setup, as seen in the previous chapter. The edge-to-edge distance between the closest signal pads of two adjacent structures is 120 µm between two columns and 248 µm between two rows. Also, the edge-to-edge distance between the closest ground pads of two adjacent structures is 172 µm between two columns and 45 µm between two rows. The signal pads have areas of 35×27 μm^2 in order to reduce pad capacitance that can be large enough at mmW frequencies to adversely impact measurements: with this pad size we have found a capacitance of approximately 6-8 fF.

Thru Each device evokes the same layout structure of the microstrip thru depicted in Fig. 3.3 and located on the floorplan at position A4. The signal trace is located in the center at M8 level and ground trace are on both sides (GSG configuration).

The ground plane is composed of a M1-to-M4 stack shunted together to avoid any electric field leaking to the lossy silicon substrate below the line (Fig. 3.3a). It is slotted to satisfy the metal density rules and is connected through vias and metals to the side ground traces. The side ground traces are within 12.1 µm to avoid coplanar waveguide modes to establish, while its distance to the ground plane is 4.9 µm. M8 (thus the line's) thickness is 3 µm, while the width of the signal trace is chosen 5.8 µm so that the characteristic impedance of this thru, as well as all the other access lines and calibration standards' lines, are set to 50 Ω. The resulting effective dielectric constant (evaluated through EM simulation) is $\varepsilon_{R,eff} \approx 3.4$.

The transmission line of the thru is divided into portions in order to define its length. The "b-b" distance is 35 µm long and does not account for the "b-c" distance which represents a non-uniformly wide transition from pad to the actual straight line (also called *access* or *launch line*). As previously discussed, after TRL calibration, the position of the reference plane is set at the center of the thru, i.e. at position "a" in Fig. 3.3b. The algorithm then allows to displace this imaginary plane to a different position by knowing the propagation constant per unit of length and using cascade matrix computation along the transmission line. To maintain the geometry properties of the line, we have arbitrarily decided to place the reference plane at position "b", thus allowing us to define the effective length of the thru for run 1 as the "b-b" distance: 35 µm. The access lines after the signal pad, i.e. the "b-c" distance, are 10.6 µm on both sides, and are the same for every test structure on the die.

Fig. 3.3c shows an isometric (side) 3D image of this passive element after being imported into an electromagnetic field simulator. We can glimpse some copper blocks around the signal pads: these unconnected metal "dummy" cubes have been added to the design of the structures to fol-

low design rules. These elements are unconnected and therefore do not affect the electrical behavior of the device.

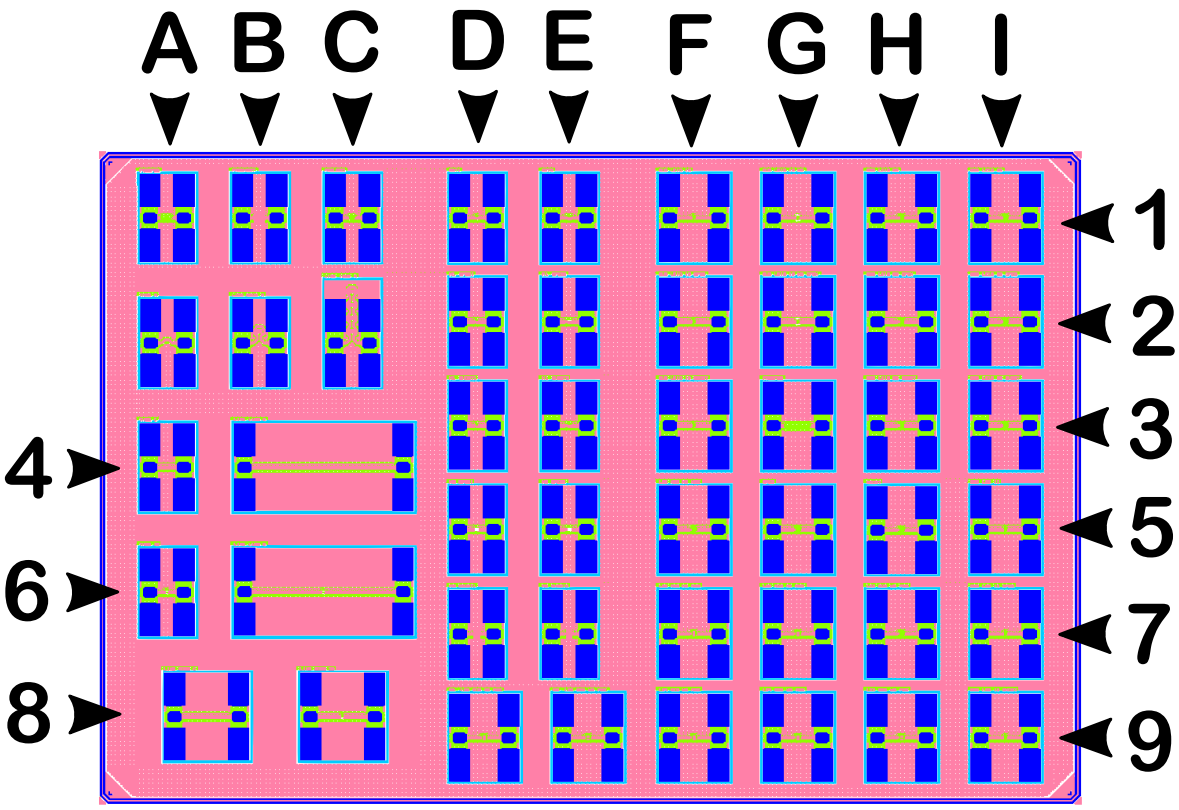


Figure 3.2: Floorplan artwork of production run 1 test structures. Columns/rows are indicated with progressive letters/numbers (see Table 3.1). Color key: pink is M8 (copper), blue are pads (aluminum), light green indicates the ground (copper), aqua is a silicon dioxide ring (SiO_2). Silicon dioxide located above M8 is not shown for clarity.

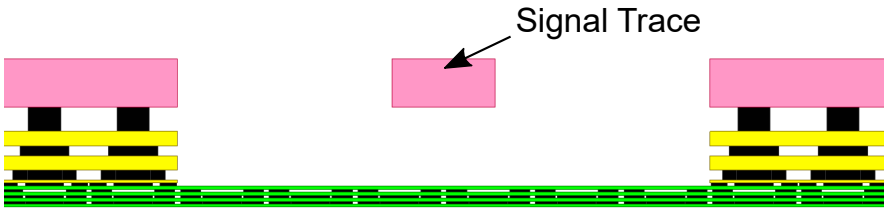
		Column Letter								
Raw Number		**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
	1	P-O	P-S	P-L	T-0 (ref.)	T-1	no use	no use	no use	no use
	2	Thru (M)	L-500G (M)	L-110G (M)	C-O (T-0)	C-O (T-1)	no use	no use	no use	no use
	3	-	-	-	C-S (T-0)	C-S (T-1)	no use	no use	no use	no use
	4	Thru	L-110G		-	-	-	-	-	-
	5	-	-	-	O-M8	O-M8	no use	no use	no use	no use
	6	Thru (3D)	L-110G (3D)		-	-	-	-	-	-
	7	-	-	-	S-M8	S-M8	no use	no use	no use	no use
	8	L-500G	-	L-500G (3D)	-	-	-	-	-	-
	9	-	-	-	no use	no use	no use	no use	no use	no use

Table 3.1: Test structure mapping for production run 1 floorplan (Fig. 3.2).

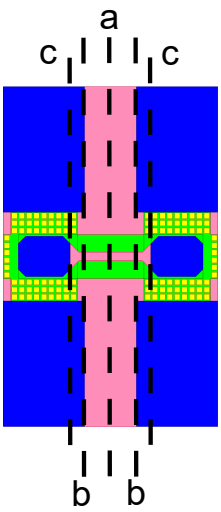
When measured over frequency, the reflection S-parameters are located in the center of the Smith chart, since no reflected wave goes back to the receiver ($\mathrm{mag}(S_{ii}) = 0 = -\infty$ dB); the transmission S-parameters start at low frequency at the extreme right of the chart, moving progressively in the capacitive region, always on the same circle ($\mathrm{mag}(S_{ij}) = 1 = 0$ dB). In a linear phase system like ours, the phase shift of the thru is linear over frequency, namely:

$$\phi = \arg(S_{ij}) = -\beta l = -\omega \tau_{\phi} \tag{3.1}$$

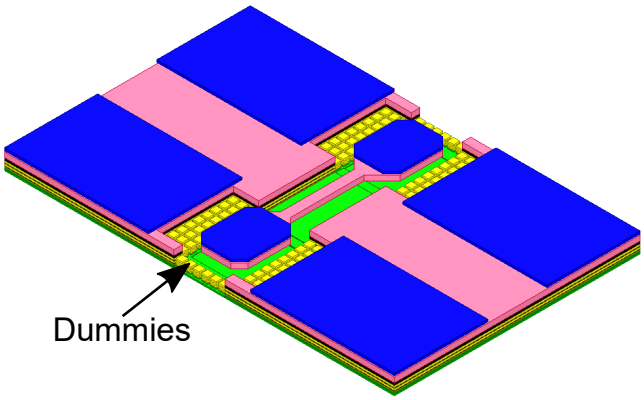
ω being the angular frequency, l the length of the thru and τ_ϕ the constant phase delay (also equal to the group delay τ_g, by $\tau_g = -\frac{d\phi}{d\omega}$). Its theoretical value, easily found by EM simulation of the intrinsic line, is around 220 fs. The resulting phase velocity is 159 m/µs.



(a) Front view without pads



(b) Top view. Reference plane is located at "b" after calibration



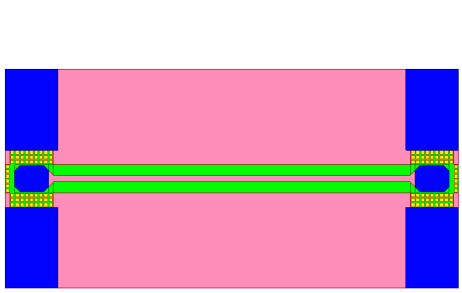
(c) Isometric view. "Dummy" unconnected metal cubes for field enclosure are indicated

Figure 3.3: Run 1 thru standard HFSS 3D model for TRL calibration. Color key: pink is M8 (copper), blue are pads (aluminum), light green indicates the M1-M4 ground (copper), yellow are layers M5-M7. Dielectric layers are not shown for clarity.

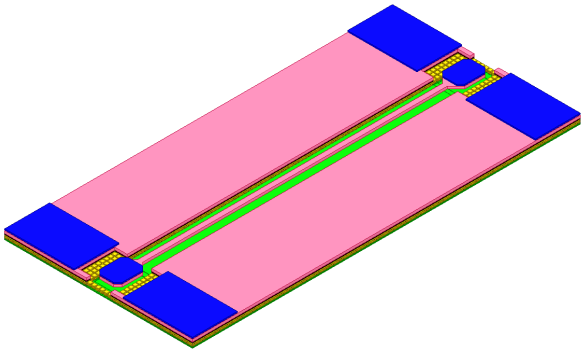
110-GHz line (L-110G) & 500-GHz line (L-500G) Following the introduction of the TRL calibration standards, we locate at position B4 the longer of the three lines, the 110-GHz line (L-110G), shown in Fig. 3.4. Its name stems from the fact that this line can be used without incurring measurement errors just within the first frequency band, namely up to 110 GHz. In fact, by referring to Eq. 2.24, we can calculate its range of validity for the TRL calibration by means of an intrinsic device simulation (more on this in the following). It yields a range of approximately 25 GHz to 200 GHz, since by taking the same conventional positions as the thru to compute its length, this turns out to be 365 µm long. Due to the different length, its area occupancy is 3.3 times more than the thru standard. At position A8, always shown in Fig. 3.4, we find the 500-GHz line (L-500G). This line is used in the upper part of the spectrum, hence its name. Its length ("b-b" distance) is in fact 115 µm, therefore its frequency validity ranges from 100 GHz up to 810 GHz. Its wafer area occupancy is again higher than the thru's, being 1.5 times more. These lines share of course the same characteristic impedance of the thru, as well as the same design.

When measured over frequency, the reflection S-parameters behave similarly to the thru. The (unwrapped) phase shift of the lines is also linear over frequency, resulting approximately in delays of 2.21 ps (L-110G) and 700 fs (L-500G).

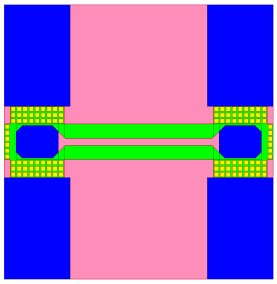
Pad-open (P-O) & pad-short (P-S) For the reflect standard, a structure called pad-open (P-O) acts as an open circuit at the RF pads' plane, thus providing high reflection for the incident waves.



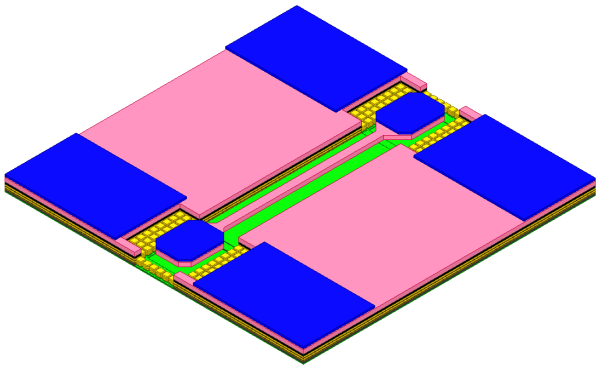
(a) Top view of 110-GHz line



(b) Isometric view of 110-GHz line



(c) Top view of 500-GHz line



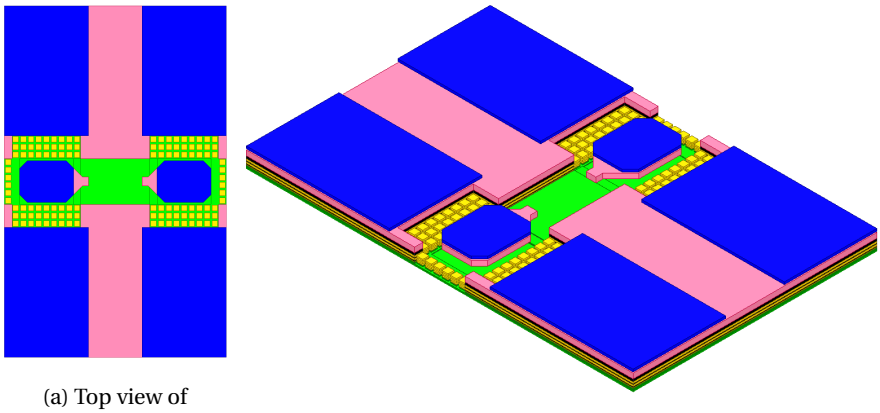
(d) Isometric view of 500-GHz line

Figure 3.4: Run 1 line standards HFSS 3D model for TRL calibration. Color key: pink is M8 (copper), blue are pads (aluminum), light green indicates the M1-M4 ground (copper), yellow are layers M5-M7. Dielectric layers are not shown for clarity.

This passive element is located at position A1 and shown in Fig. 3.5. The symmetrical open circuit is created at the edge of the access lines (position b of the thru, Fig. 3.3b), thus the open-edge-to-open-edge distance is the same as the thru length, 35 µm. Its dual structure, called pad-short (P-S) acts as a short circuit built at pad level, it is located at position B1 and shown in the same figure. At the edge of the access line, a choke-point connects the pads through a metal extension directly to the side ground pads. In the following, P-O will be chosen as the main solution for the reflect standard in the TRL calibration, except when calibrating a transistor short: in that case, P-S will be preferred [143].

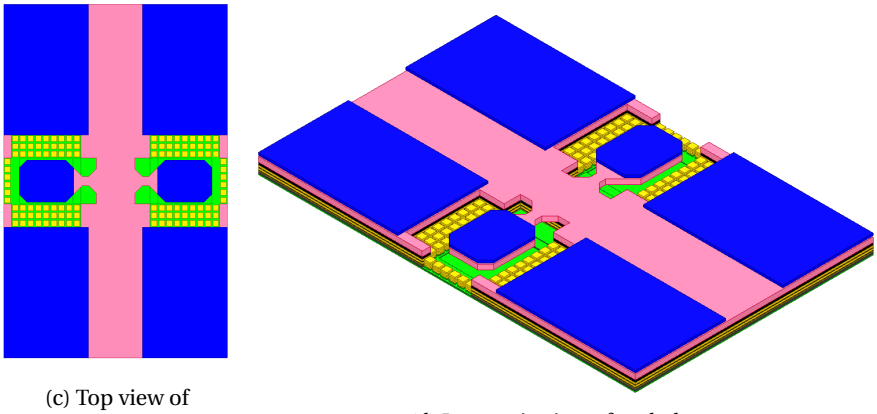
When measured over frequency, the reflection S-parameters are located in the right end (P-O) and left end (P-S) of the Smith chart, since the reflected wave bounces back to the generating port (mag(S_{ii}) = 1 = 0 dB), progressively moving in the capacitive (P-O) or inductive (P-S) region, always on the same circle; the transmission S-parameters, on the other hand, are located in the center of the chart, since no signal is detected at the opposite port (mag(S_{ij}) = 0 = $-\infty$ dB). By a simple evaluation of the DC current at port 1 over the applied voltage, we can find the input resistance: $R_s = V_1/I_1$. This resistance represents the series contribution of all the connections (cables, probe, etc...) from the bias generator to the ground, which is located, in the case of the pad-short, at the probe tip. This resistance is slightly dependent on the contact quality, but also on the probes employed. We experienced values ranging from 1.5 to 2 Ω.

Pad-load (P-L) Moreover, to perform an impedance correction, one extra pad-load (P-L) structure consisting of four loads of approximately 100 Ω each (positioned in shunt pairs, for a resulting 50 Ω at each port), located below the M1 level, is considered. The access line has the same layout design as the thru and line standards. It is located at position C1 and shown in Fig. 3.6. The principle of impedance correction will be detailed in the following. The BEOL consists in the same



(a) Top view of pad-open

(b) Isometric view of pad-open



(c) Top view of pad-short

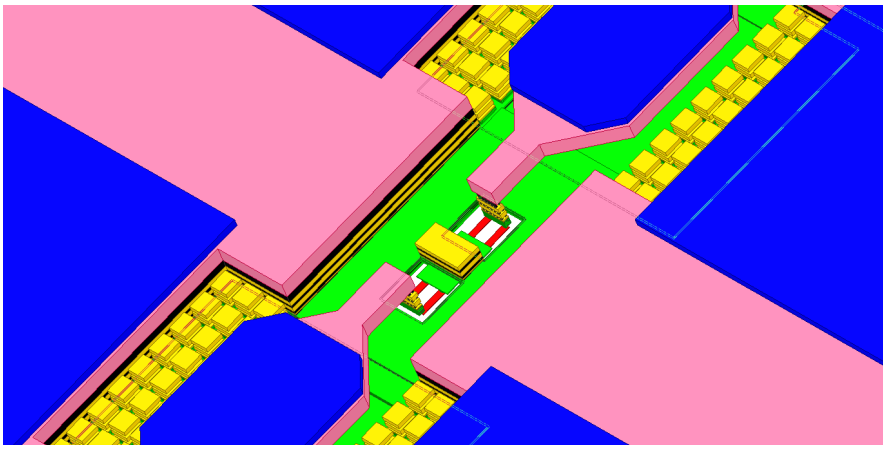
(d) Isometric view of pad-short

Figure 3.5: Run 1 reflect standards HFSS 3D model for TRL calibration. Color key: pink is M8 (copper), blue are pads (aluminum), light green indicates the M1-M4 ground (copper), yellow are layers M5-M7. Dielectric layers are not shown for clarity.

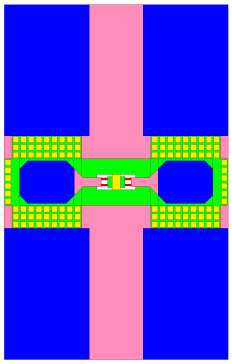
sequence of metal and via connections, from pad level (M8) to transistor level (M1), as the active device. M1 is contacted to the polysilicon resistances, two for each port (see Fig. 3.6a), measuring approximately 36 Ω at port 1 and 2: the electrical conductivity of each load is easily computed by knowledge of the geometrical properties of these resistances and yields 0.65 S/µm. We assume here that the fabricated load is real and equal to its DC value (the DC contribution of connections has been removed from the actual measured value): the difference from this ideal behavior and from the target 50 Ω stems from the difficulty to fabricate in such complex technologies a load with stable and controlled properties, particularly as frequency rises.

Over frequency, the reflection as well as the transmission S-parameters are located approximately in the center of the Smith chart: in transmission, the input signal is not detected by the opposite port, since any direct connection is absent; in reflection, the input signal is (theoretically) absorbed by the 50 Ω load, and again, nothing is detected. As we have seen, however, this is just an ideal scenario, since the designed capacitance can hardly comply with this specification. Part of the signal is therefore reflected and the Smith chart locus of S_{ii} is somewhere in the proximity of the center. By evaluating the input resistance R_i using the measured values of the input current, and by removing the contribution of the series resistance of the connections retrieved thanks to the P-O, R_s, we can find the designed load resistances: this is how we computed the approximate value of 36 Ω.

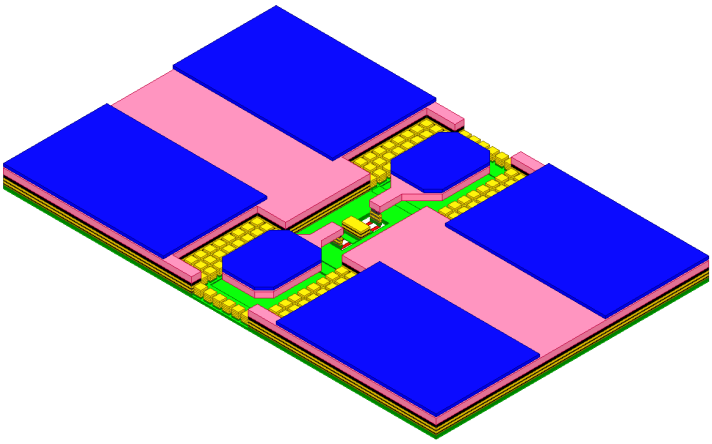
Complete-open (C-O) & complete-short (C-S) mmW technologies need to account for BEOL electrical parasitic effects, e.g. the capacitances due to the base, emitter and collector accesses, which hide the junction capacitances of the HBT, or the inductances in long access+BEOL connections. Hence, in addition to the on-wafer TRL calibration kit structures, two test structures



(a) Zoom on the polysilicon loads below M1



(b) Top view of pad-load



(c) Isometric view of pad-load

Figure 3.6: Run 1 load standard HFSS 3D model for impedance correction. Color key: pink is M8 (copper), blue are pads (aluminum), light green indicates the M1-M4 ground (copper), yellow are layers M5-M7, red are the loads (polysilicon). Dielectric layers are not shown for clarity.

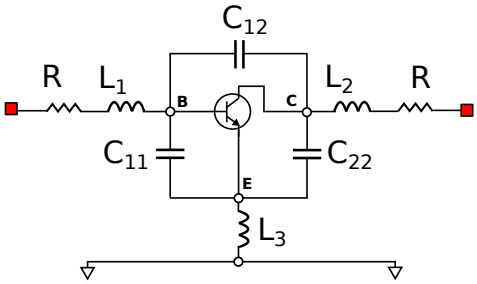
are dedicated to the de-embedding of the transistor accesses. These two test structures are called complete-open (C-O, at position D2 and E2) and complete-short (C-S, at position D3 and E3) and, as their names suggest, they provide an open/short circuit down at the M1 level, with the same access BEOL design as the transistor's (Fig. 3.1).

In Fig. 3.7 we consider again the equivalent circuit for the short-open de-embedding as presented in Fig. 2.10, where the impedances/admittances are simply displayed as inductances /capacitances (Fig. 3.7a), and we also show the location of these parasitic elements on the 3D model of the BEOL itself (Fig. 3.7b). Note that due to the lumped nature of our simple model, this representation inevitably falls short in accuracy. These de-embedding standards can therefore be used to remove the capacitances due to the BEOL connections linking the access lines to the transistor level; i.e. the distributed capacitance formed between the connections of port 1 (leading to the HBT's base) and ground (leading to the HBT's emitter), here called C_{11}, the distributed capacitance formed between the connections of port 2 (leading to the HBT's collector) and ground, called C_{22} and one formed between the connections of port 1 and 2, called coupling capacitance C_{12}, as well as all the corresponding inductances: L_1, L_2, and the mutual inductance L_3. In order to compute these values, the imaginary parts (divided by ω) of Eq. 2.25 are taken. The series resistances are also removed by this approach.

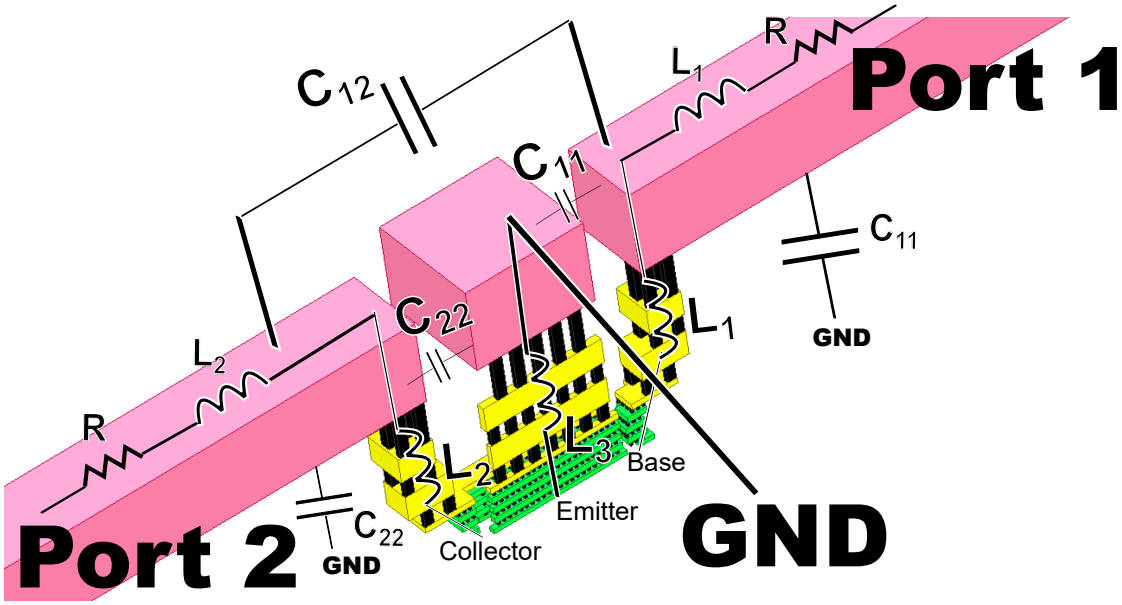
Fig. 3.8 gives a detailed glimpse to the differences in the BEOL between these two structures and the reference transistor T-0 (located at position D1 and whose emitter drawn size is 0.2×5 µm). The fact that C-O does not have any contact pins to the underlying substrate may produce some artifacts in the de-embedding process (over-de-embedding, hence an over-estimated HBT perfor-

mance) because of the unwanted capacitance between the bottom metal M1 and the substrate. Similarly to P-S, the ratio between the DC current at port 1 and the voltage will provide a series resistance which is once again linked to the contribution of all the connections but also of the BEOL, since the ground is now located at M1. We can find therefore, by substraction, this resistance, which turns out to be smaller compared to the other (less than 1 Ω).

Moreover, C-O and C-S are representative because of their elaborate BEOL and importance in the transistor measurements, and since they are not involved in first tier of calibration, they will be often chosen as a non-active DUTs for calibration verification, of measurements and simulations.



(a) Two-step circuit model for short-open de-embedding of the HBT



(b) Side view of the HBT's BEOL visualising every parasitic circuit element (GND connection is partially hidden)

Figure 3.7: Run 1 equivalent circuit for transistor de-embedding.

Other test structures As we can see in Table 3.1, some structures have not be described so far and will be quickly introduced now for a subsequent adoption.

- The test structures marked by "(M)" represent the "meander" lines: these lines have been designed to be used instead of regular straight line in the TRL calibration, they have an effective length similar to the corresponding straight lines but they occupy the same area of a thru;
- The test structures marked by "(3D)" represent the "3D" lines: these lines aim to take into account also the metal stacks of the BEOL and are designed for a single-tier TRL calibration;
- O-M8 and S-M8 are open/short circuits created at M8 level, far from the pads, at a position closer to "a" in Fig. 3.3b. They are designed for being the reflect standards in a TRL calibration, as an alternative to P-O and P-S;

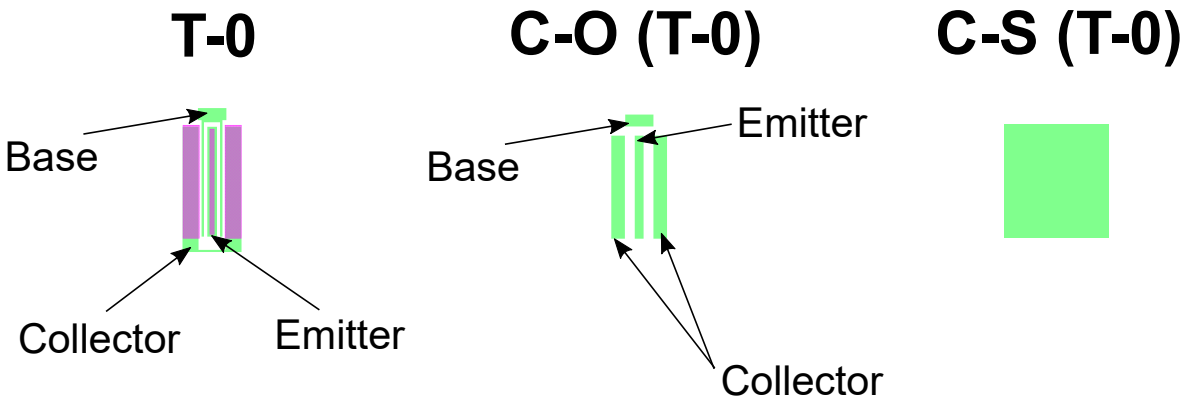


Figure 3.8: Top view of run 1 reference HBT T-0 and its C-O & C-S. Light green color is M1, purple are the underlying active regions (collector and emitter), connected to M1 by contact pins (not shown).

- T-1 represents a transistor with different geometry, being its emitter length double the length of the reference transistor T-0: $A_E^{T1} = 0.2 \times 10\ \mu m^2$. It is present for modelling purposes and won't be exploited in this work.

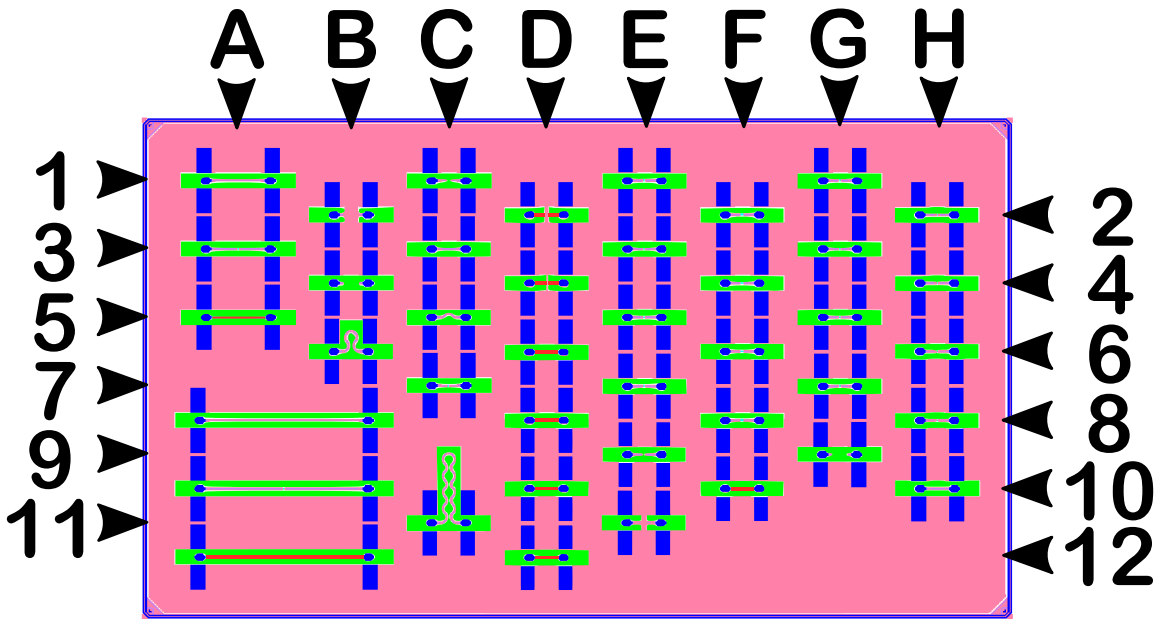


Figure 3.9: Floorplan artwork of production run 2 test structures. Columns/rows are indicated with progressive letters/numbers (see Table 3.2). Color key: pink is M8 (copper), blue are pads (aluminum), light green indicates the ground (copper), red is M3 (copper). Silicon dioxide located above M8 is not shown for clarity. M8 extends uninterruptly to all the surface and is connected directly to the ground plane at M1 (continuous ground plane).

3.1.2 Production Run 2

The test structures in this novel mask are largely based on run 1's, but their layout has been changed (see Fig. 3.9 and structure mapping on Table 3.2). The first most notable difference is the removal of the SiO_2 "ring", which is a controversial design element that possibly stems undesired EM interactions and yet, as already pointed out, it is commonly employed, in certain contexts, with no particular care taken into isolating the underlying substrate from the region of probe excitation. The ground plane, which is now composed of M1 only (170 nm thick, in green in Fig. 3.11a), instead of the M1-M4 stack (1300 nm thick), is therefore expanded to the whole die, and is common to all the devices. We decided to reduce the thickness of the ground since we haven't experienced

		Column Letter							
		A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Raw Number	**1**	L-500G	-	P-L	-	T-0 (ref.)	-	T-2	-
	2	-	P-S	-	P-S (M3)	-	T-1	-	T-3
	3	L-500G (3D)	-	Thru	-	C-O (T-0)	-	C-O (T-2)	-
	4	-	P-O	-	P-O (M3)	-	C-O (T-1)	-	C-O (T-3)
	5	L-500G (M3)	-	Thru (M)	-	C-O (no-T0)	-	C-O (no-T2)	-
	6	-	L-500G (M)	-	P-L (M3)	-	C-O (no-T1)	-	C-O (no-T3)
	7	-	-	Thru (3D)	-	C-S (T-0)	-	C-S (T-2)	-
	8	L-110G		-	Thru (M3)	-	C-S (T-1)	-	C-S (T-3)
	9	-	-	-	-	O-M8	-	no use	-
	10	L-110G (3D)		-	C-O (no-TM3)	-	T-M3	-	no use
	11	-	-	L-110G (M)	-	S-M8	-	-	-
	12	L-110G (M3)		-	C-S (T-M3)	-	-	-	-

Table 3.2: Test structure mapping for production run 2 floorplan (Fig. 3.9).

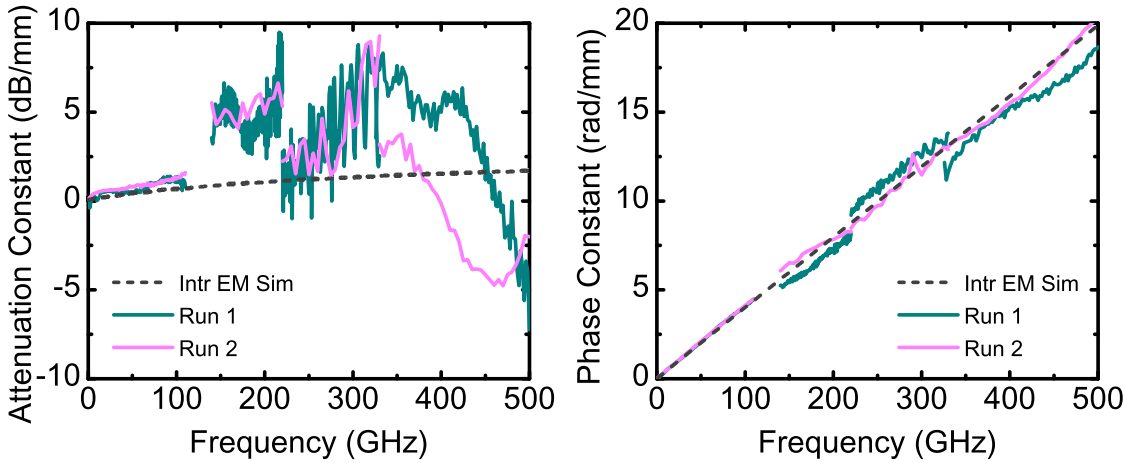


Figure 3.10: Propagation constant comparison between run 1 and 2 (same probes in each range are used). Higher number of points on run 1 may give the appearance of a less sharp trend.

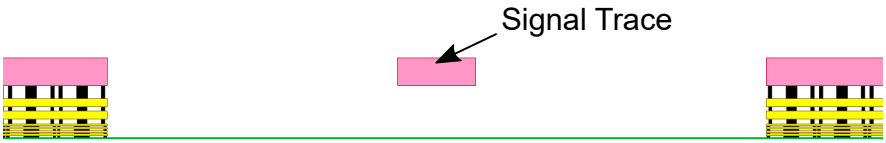
any considerable field penetration underneath, even though it is true that the skin depth at 20 GHz is 460 nm and that M1 is just 170 nm, but the possibility for any energy leak in the lossy silicon substrate is very low and restricted in frequency. According to preliminary simulation models, we did not observe any consistent difference in the propagation constant, either. As for measurements, we can see a direct comparison of the measured attenuation and phase constants on both runs in Fig. 3.10 and confirm that curves are quite alike (very similar shapes), even though the linearity of β has been improved on run 2. More insight on these values will be given next. The absence of slots and the extension of the ground plane together result in an uninterrupted sequence of M1-to-M8 ground volumes connecting all the side grounds together via the M1 plane and is seen from M8 as a boundless plane.

On the other hand, pads are still compatible to both 50 and 100 µm pitch probes, but their design has also deeply changed. Following the considerations of Yadav *et al.* [142] on early run 1 measurements, we have drawn inspiration by Seelmann-Eggebert *et al.* [107] for a new RF pad design. They observed that, since parasitic modes originate at the contact point of the probe with the substrate, it would be recommended to modify the design in the vicinity of the tips. They suggest using a grounded guard ring preventing waves to escape to the side opposite to the transmission line. We decided to implement this shielding structure behind our pads, which is therefore forming a sort of cage for the signal, since the boundless ground extends from bottom to top of the metal stack (Fig. 3.11b). The combination of this two design elements has been dubbed "continuous ground plane" to stress on its ubiquity among the structures on the die, its intent being to reduce the probe-to-substrate EM coupling and coupling with neighbors.

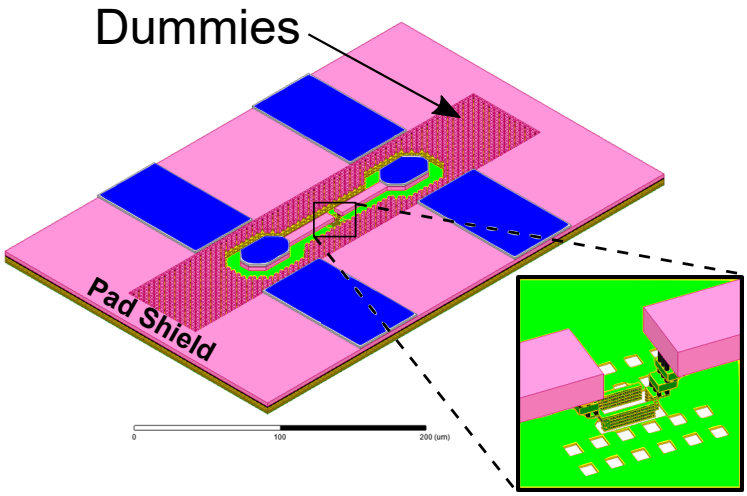
From the direct comparison of the artworks of the two masks (compare Fig. 3.2, 3.9), an-

other fundamental difference can be immediately spotted. On run 1, the structures are aligned in columns and rows in a matrix format. Structures on run 2, however, are placed differently: a particular "chessboard" layout with staggered rows is used in order to further reduce EM coupling between the RF probes and the structures adjacent to the DUT, i.e. its "neighbors".

Since run 2 replicates most of the properties already presented for run 1, we decided to list the main size and physical differences in Table 3.3. Below are presented the calibration kit and devices on the mask.



(a) Thru front view without pads



(b) Complete open isometric view. "Dummy" unconnected metal cubes for field enclosure and pad "shielding" structure of the continuous ground plane are indicated. Inset shows the BEOL of the C-O

Figure 3.11: Run 2 HFSS 3D model of thru and C-O. Color key: pink is M8 (copper), blue are pads (aluminum), light green indicates the M1 ground (copper), yellow are layers M2-M7. Dielectric layers are not shown for clarity.

Standard calibration kit The standard calibration kit presented for run 1 has been redesigned with the appropriate layout changes and different geometrical dimensions (see Fig. 3.12). The thru standard (located at position C3 in Fig. 3.9, see Table 3.2) is 65 µm long now and again, its length does not take into account the access lines (the usual pad-open edge to pad-open edge is considered). The 110-GHz line (located at position A8) and 500-GHz line (position A1) are 595 and 185 µm, respectively. Pad-open (position B4) and pad-short (position B2) are also akin. Pad-load (position C1) is now made of two (instead of four) polysilicon loads of 50 Ω each, but the DC measured value at both ports is again around 36 Ω. The BEOL topology has also been changed with respect to run 1. It accounts for the different M1 footprints in Fig. 3.14: whilst previously the connection emitter-ground was made through a BEOL stack ascending to M8 then descending back to the ground plane, it is now at M1, by connecting the emitter contact directly to the ground plane, without any stack (Fig. 3.13). This is clearly visible in the T-0 and C-O of Fig. 3.14: the emitter central trace is connected to the surrounding ground plane (partially not shown); the base and collector traces have also different shapes. As the C-O of run 1 has a similar footprint to the complete-open "without transistor" (no-T0) of run 2, it will be considered as the main verification standard in the following. We have also added in run 2 a different version of complete-open, simply called C-O, where the emitter/base/collector contacts are gone, but the emitter/base/collector

	Property	Run 1	Run 2
General	SiO_2 ring or slot	Yes	No
	Chessboard layout	No	Yes
	Ground plane	M1 to M4	M1
	Probe compatibility	50 & 100 µm pitch	
		Value (µm)	
Pad geometry	Signal pad column distance	120	210
	Signal pad row distance	248	245
	Ground pad column distance	172	180
	Ground pad row distance	45	15
	Signal pad area	35 × 27	40 × 25
Transmission line geometry	S-G horizontal distance	12.1	28.6
	S-G vertical distance	4.9	5.6
	Ground plane thickness	1.3	0.17
	Line thickness	3	
	Line width	5.8	7.7
Line lengths	Access length	10.6	15
	Thru length	35	65
	L-110G length	365	595
	L-500G length	115	185
		Value	
Microstrip	Characteristic impedance	50 Ω	
	Effective dielectric constant	3.4	3.6
Line delays	Thru phase delay	220 fs	420 fs
	L-110G phase delay	2210 fs	3880 fs
	L-500G phase delay	700 fs	1180 fs
Line validity	L-110G validity range	25-200 GHz	14-120 GHz
	L-500G validity range	100-810 GHz	65-520 GHz
Wafer occupancy	L-110 area excess (w.r.t. thru)	940%	815%
	L-500 area excess (w.r.t. thru)	230%	185%
DC characteristics	P-S series resistance	1.5-2 Ω	
	P-L load resistance	36 Ω	
	P-L load conductivity	0.65 S/µm	

Table 3.3: List of topological, geometrical, physical and electrical properties and their values in run 1 & 2.

M1 footprints are present, besides, in particular, the transistor underneath, recreating the environment of the transistor's BEOL in the most accurate way.

On top of Fig. 3.15 we show the different values of the capacitances of these two configurations of C-O (C_{22} is not shown for clarity). The dotted traces represent the intrinsic (reference) value, the scattered points the (TRL-calibrated) measurement. The capacitance values of C-O are consequently higher than C-O (no-T0) since they embed the contributions from the metal pins in M1, particularly at HF. The rise of the coupling capacitance above 300 GHz follows the intrinsic simulation, as expected. This few hundreds attofarad difference has a beneficial effect on the transistor measurements, and in particular in retrieving the transit frequency f_T (which is dependent on the equivalent base-emitter capacitance) and maximum oscillation frequency f_{max} (which is dependent on f_T itself and the equivalent base-collector capacitance): on average, f_T data are 2% closer to simulation, while f_{max} data are 0.3% closer with just this change on the de-embedding step.

Other test structures Here are presented special test structures not present in a standard calibration kit or for device test, specific to run 2:

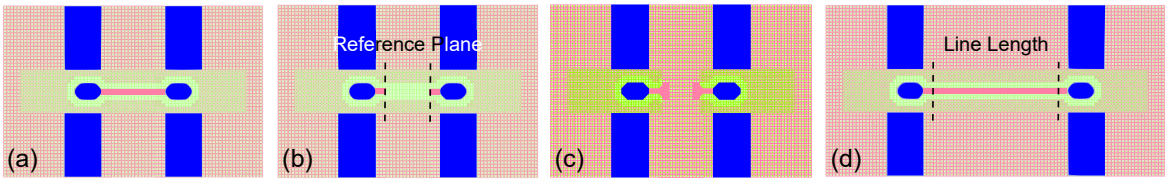


Figure 3.12: Artwork of the on-wafer TRL calibration kit. (a) thru, (b) P-O, (c) P-S, (d) L-500G. In pad-open, post-calibration reference planes are shown with black dashed line.

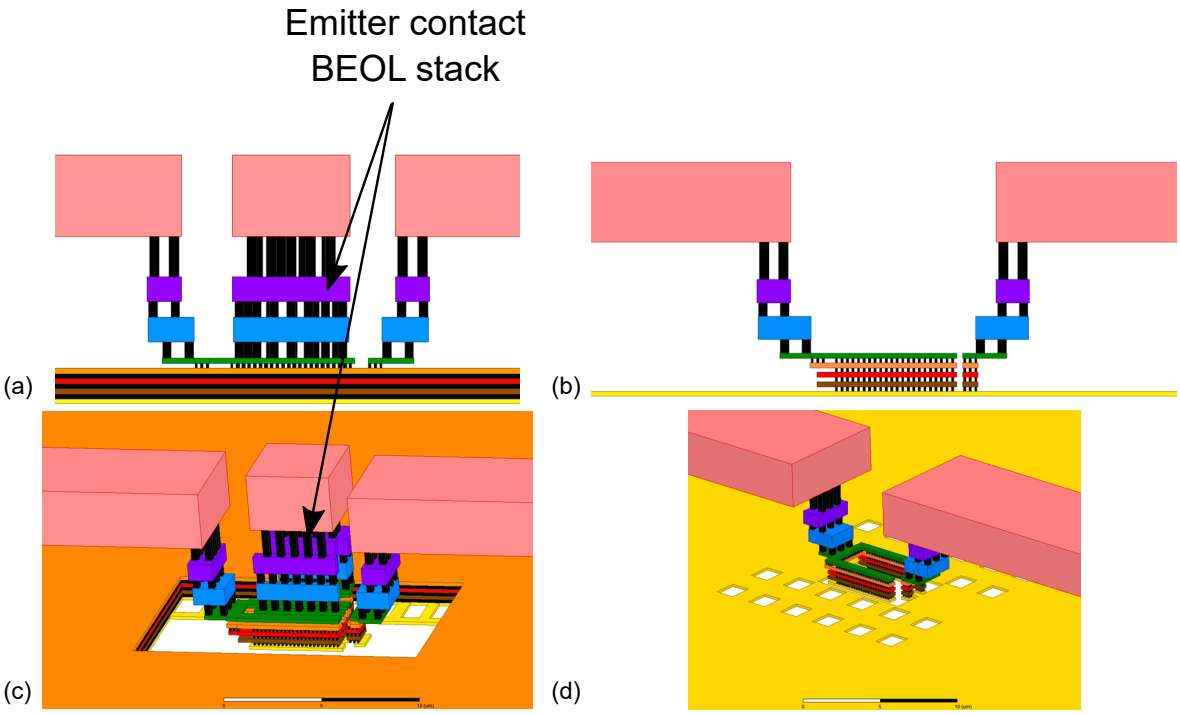


Figure 3.13: Complete-open. Cross section details from run 1 (a) and run 2 (b) show metal layers with different colors from M1 (yellow) to M8 (pink) in color, while vias are in black. Side 3D details from run 1 (c) and run 2 (d) are also shown. On run 1, the metal stack connection bridges the emitter contact (not visible due to perspective) to the M1-M4 ground plane.

- The test structures marked by "(M3)" are a variant of standard structures where the access lines are located at metal 3 instead of metal 8. A specific transistor as well as its associated de-embedding standards and a dedicated calibration kit have been designed;
- T-2 and T-3, with associated C-O and C-S, represent transistors with different geometry, as their emitter widths change: $A_E^{T2} = 0.3 \times 5\ \mu m^2$, $A_E^{T3} = 0.42 \times 5\ \mu m^2$. They are present for modelling purposes and won't be studied here.

3.1.3 Comparison to Other Implementations

Recent works used a variety of different solutions to implement the layout of the structures; it is interesting to compare the dimension of some of the transmission lines and some of the layout properties. In [131], Williams *et al.* designed the microstrip thru, 400 µm long, on low-loss single-mode (up to 750 GHz) bisbenzocyclobutene (BCB) monomers and put structures 150 µm apart. In [130, 132], Williams *et al.* used a 300 µm long CPW thru instead, laying on a dielectric with relative constant of about 4. At TU Delft, Galatro and Spirito [39] built a CPW transmission line on a silicon dioxide substrate and gave the thru a length of 100 µm (pads excluded), with a larger width to minimize losses. Fregonese *et al.* extremely reduced the length of the thru in [32] and took a microstrip approach, just like Phung *et al.* in [79] (we consider only the microstrip layout), with a BCB substrate and no side ground, which have made a study of aligned structures' distances

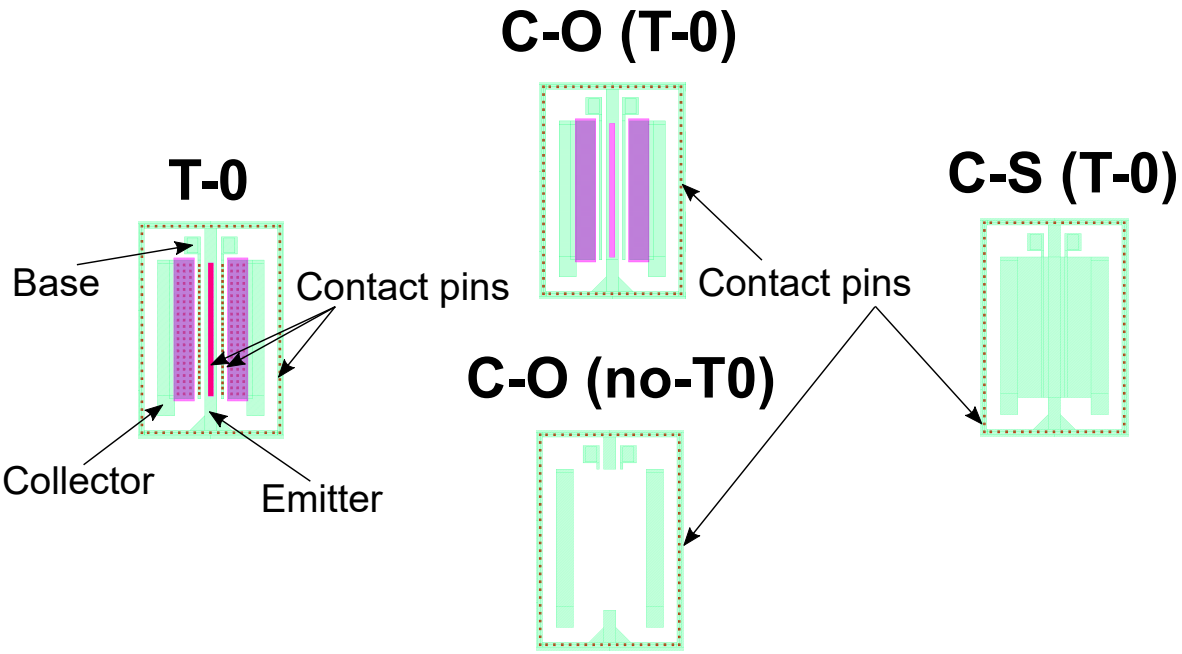


Figure 3.14: Top view of run 2 reference HBT T-0 and its C-O & C-S. Light green color is M1, purple are the underlying active regions (collector and emitter), connected to M1 by contact pins (in red).

to evaluate neighboring coupling.

Table 3.4 makes a direct comparison of the previous works' implementations. It is worth to mention that while we do not have any information on the type of layout topology for the first three presented works, we can guess that their layout is a "classic" aligned one, different from the chessboard topology adopted in run 2, as no mention about the devices' placement is provided. We highlight that the inter-structure distance displayed for run 2 is achieved thanks to the chessboard layout, while preserving a short "in-line" structure distance (210 µm). The thru length of run 1 and 2, calculated by considering the "b-b" distance, like the other authors did, has been increased in run 2 and represents a good compromise between large structure wafer occupancy (fabrication costs) and physical constrains (possible cross-talk in smaller structures, as evidenced in [32]).

While the coupling with neighbors will be studied separately further on, let us mention some recommendations provided by Williams *et al.* in an extensive study on crosstalk (on CPW structures) [133]. In this paper, the line on a CPW topology is located in different substrates (a thick ceramic substrate, a thin ceramic substrate placed on quartz or metal bulks) and different crosstalk standards, used to retrieve the crosstalk contribution with a 16-term model, with various geometrical features, are employed. In our study, no such structures are present, since we employ a 8-term correction that does not take care directly of the crosstalk: designing those structures would make the calibration non-trivial and increase wafer occupancy. However, complete-open de-embedding standard partially removes the port 1 to port 2 coupling between the probes and the substrate coupling. In CPW topology, higher-order modes are excited early with frequency, and they cause strong ripples and resonances, particularly if a metal chuck is employed [79] (Fig. 3.16a). As stated by Williams *et al.*, "coupling to parallel-plate modes [...] increases when the substrate thickness decreases" [133] and crosstalk is also hardly corrected by the more complete 16-terms model "possibly because these modes lead to a violation of the assumption of a single-mode port at the on-wafer reference plane" [133]. Uncorrected crosstalk has a very clear effect, according to the authors, i.e. "gives rise to large ripples [...] due to parallel-plate and substrate modes" [133] on the measurements (of the scattering parameters). This signature (ripples for crosstalk) is common to crosstalk also in microstrip lines, since it's independent of the planar topology chosen (Fig. 3.16b). Eventually, the general recommendations provided by the authors that have been taken in this work are: 1) a small pitch and pad area, to minimize coupling below the pads, 2)

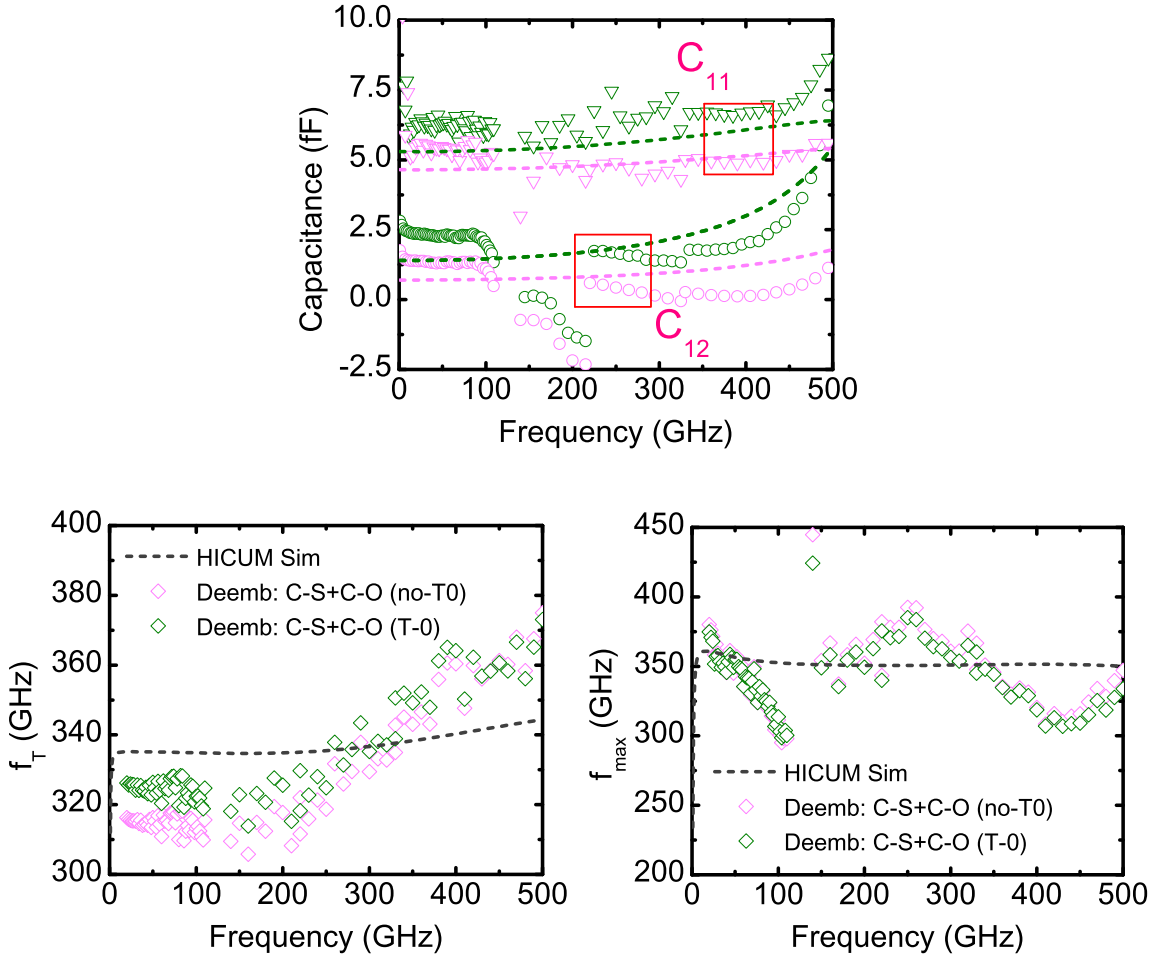
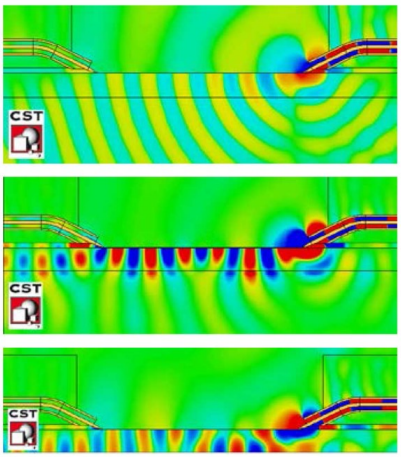


Figure 3.15: Capacitance measured and intrinsic values of both complete opens. Color key: green is from C-O and pink is from C-O (no-T0). The two complete open structures have been subsequently applied on the de-embedding to retrieve the transit frequency and maximum oscillation frequency of the HBT measured at $V_{CB} = 0\,V$, $V_{BE} = 0.9\,V$.

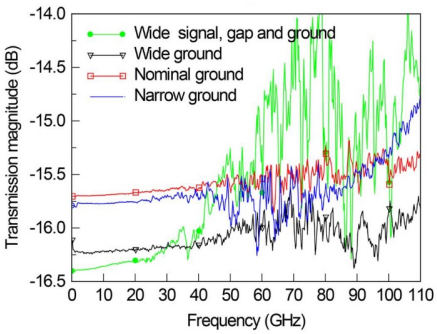
small width of the central conductor, fine-turned to avoid large conductor losses, in order to reduce radiation and multiple propagation modes, 3) short access lines.

Our line topology is different, though: silicon technologies employ complex BEOL with 7 or more metal layers and higher-mode generating CPW (widely employed for III-V technologies) can be discarded to opt for a more controlled microstrip instead. A microstrip mainly supports a quasi-TEM mode but higher-order modes can be generated, at very HF. For example, a TM or TE surface wave mode (when discontinuities are present, TE surface modes are important) are created and the quasi-TEM mode can couple with it: the frequencies where these couplings appear are high [84]. Another example are parallel-plate waveguide modes between the strip and the ground plane, which, for thin strips and therefore high fringing fields, might occur on a larger surface; again, the low transmission line thickness and relative permittivity of our lines avoid these types of mode to propagate [84].

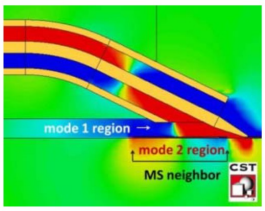
The transmission coefficient for higher-order modes originated in our run 2 thru are shown in Fig. 3.17. As we can see, these modes do not propagate as intensely as the main mode, and effectively do not exist until very high frequencies. More in general, Phung *et al.* discussed in [77] that three types of parasitic modes can be identified between a multilayer thin-film microstrip technology like ours and a probe, generated in proximity of the tips: 1) fields between the outer material of the probe and the metallization of adjacent structures ("mode 1" in Fig. 3.16c), 2) fields between the ground metallization and the chuck (into the substrate, similarly to a parallel-plate mode: "mode 2" in Fig. 3.16c), 3) fields between the probe and the backside metallization. While



(a) From [133]



(b) From [133]



(c) From [77]

Figure 3.16: (a) Field plots simulations with the CPW substrate placed on: ceramic (top), quartz (middle), and metal (bottom). (b) Ripples caused by to uncorrected crosstalk due to parallel-plate and substrate modes on a CPW, when the substrate is placed on a metal chuck. (c) E-field in the longitudinal cross section with a short microstrip neighbor below the probe (blue color means negative, red color positive and green color zero field values): two modes are highlighted.

Property	**NIST [131]**	**NIST [130]**	**TU Delft [39]**	**U. Bdx [32]**	**FBH [79]**	**Run 1 (this work)**	**Run 2 (this work)**
Chessboard layout	N/A	N/A	N/A	No	No	No	Yes
TL topology	MS	CPW	CPW	MS	MS	MS	MS
Dielectric height (µm)	8	6.3	8.8	4	16	4.9	5.6
S-G horizontal distance (µm)	N/A	≈ 4	10	28	-	12.1	28.6
TL thickness (µm)	N/A	≈ 3	3	2.8	11	3	3
TL width (µm)	22	6	30	5	37	5.8	7.7
Thru length (µm)	400	300	100	50	900	55	95
Effective dielectric constant	2.6	3.5	4.1 (rel.)	N/A	2.6	3.4	3.6
Inter-probe distance (est.) (µm)	N/A	350	250	90	N/A	90	135
Inter-structure distance (µm)	150	N/A	N/A	24	100 to 1000	275	322
Signal pad area (μm^2)	N/A	40 × 30	50 × 30	38 × 38	N/A	35 × 27	40 × 25

Table 3.4: Comparison of on-wafer structures' design from various authors.

our run 1, because of its openings into the substrate, can let pass all three types of coupling, run 2 with its boundless plane suppresses completely the latter two. The only existing coupling mode can be that below the probe to the homogeneous ground metal volume.

Therefore, layout plays a fundamental role into reducing the coupling, which, as we have seen, even by highly elaborated calibration error models cannot fully be corrected. For this reason, we increased the inter-probe distance in run 2 (Table 3.4) while keeping small access lines and a relatively small thru through an accurate design of pads and access (the estimation of the inter-probe distance is, in fact, done by considering the "b-b" thru length and a constant portion of the pad length), in an attempt to trade-off between reduced losses and suppression of spurious modes.

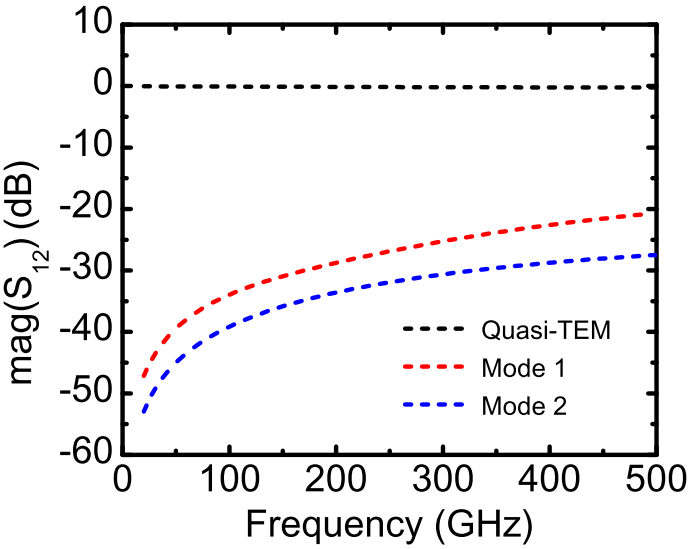


Figure 3.17: Run 2 thru transmission of higher order modes.

3.2 Simulation Setup

We have briefly introduced in the previous section the electromagnetic simulation for validation of measurements on passive devices. Electromagnetic field solvers are specialized programs capable of solving Maxwell's equations directly. They represent a branch of electronic design automation (EDA) and model the interactions of EM fields with the environment and physical objects. EM field solvers provide solutions to real-world problems that are not analytically calculable. From the simulated electric (E) and magnetic (H) fields, S-parameters can be computed much as during a real-world measurement if the DUT and the measurement set are accurately replicated in the simulator. Also, EM software tools are able to display meshed field overlays that help designers visualize EM field radiation and interactions. However, while passive layouts obey to classic physics rules, active components such as transistors and varactors underlie semiconductor physical effects such as drift-diffusion equations for carrier transport, and it is not possible to directly simulate their responses. Throughout this work we will make use of Ansys High-Frequency Structure Simulator (HFSS).

Fig. 3.18 presents the flowchart for simulating (and calibrating) a single test structure, either active or passive. Three distinct scenarios will be studied: simulation of the intrinsic device, simulation of the complete device (RF pads and access line) with RF probe tip models, probe simulation with lumped port to integrate the HICUM transistor model to replicate the real-world measurement environment of an active device.

3.2.1 Intrinsic Electromagnetic Simulation

To study the intrinsic calibration standards properties and to verify them, we first take the layout 2D data or GDSII file (Graphical Data Stream Information Interchange) that represents planar geometric shapes and other information on the test structures in a hierarchical form. The meshed data of each layer need to be properly simplified wherever notched contours or small holes (smaller than the minimum signal wavelength, the holes' size being approximately $1.8 \times 1.8\ \mu m^2$) are present. We also take care of removing all the dummies since they are not electrically connected to the structures and their presence considerably increases the simulation time. The approach of simplification of the different layers of dummies shown in Fig. 3.19, demonstrates that no variation in the main L-500G parameters occurs by keeping the lower metal dummies, too. However, slightly higher losses and an overestimation of the transmitted signal are reached with an even simpler model. The increased $|S_{21}|$ (by less than 0.1 dB) and α (by 0.2 dB/mm) are perfectly

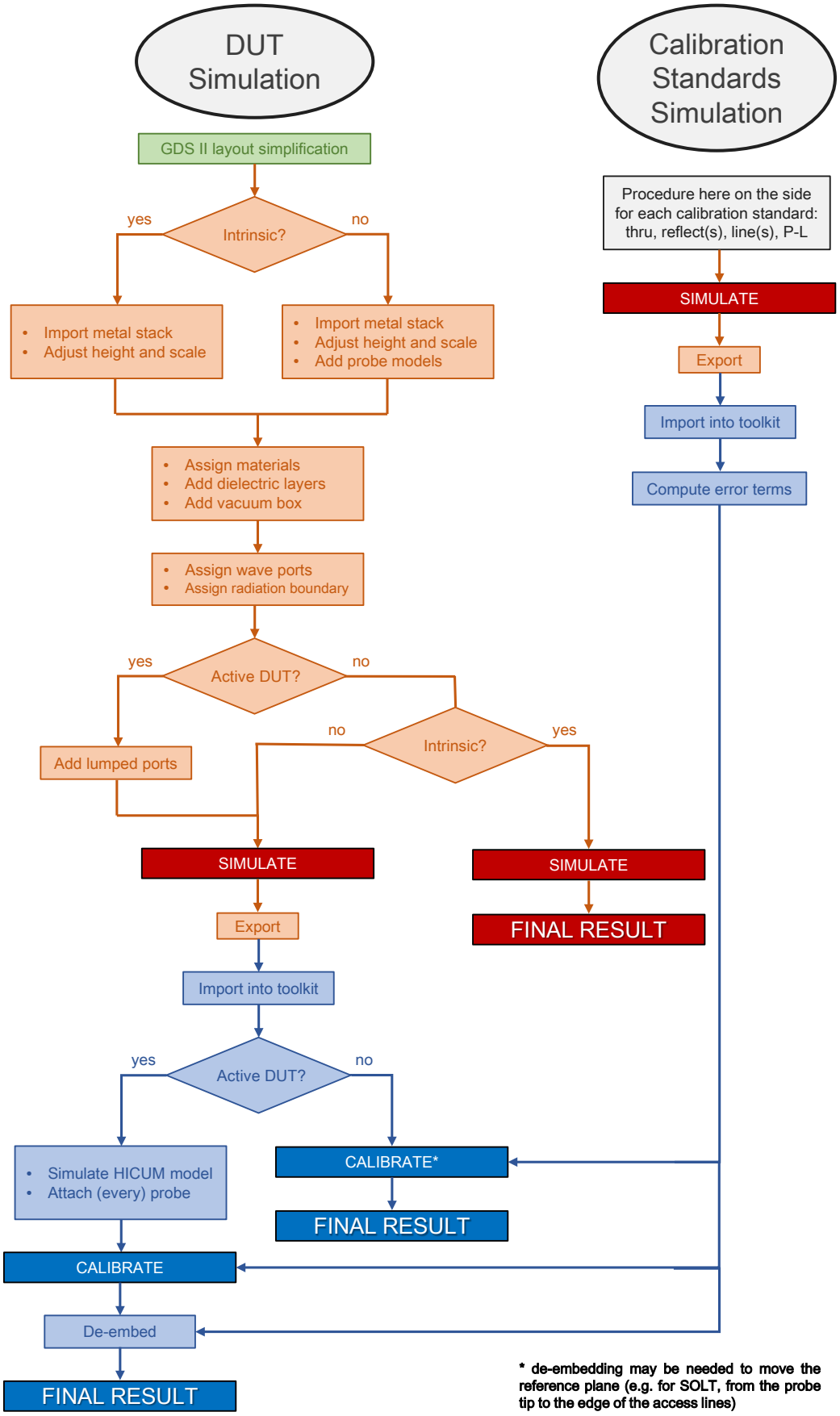


Figure 3.18: Flowchart of the various single structure simulation scenarios. Color key: green part takes place in a layout editor, orange in HFSS, blue in IC-CAP.

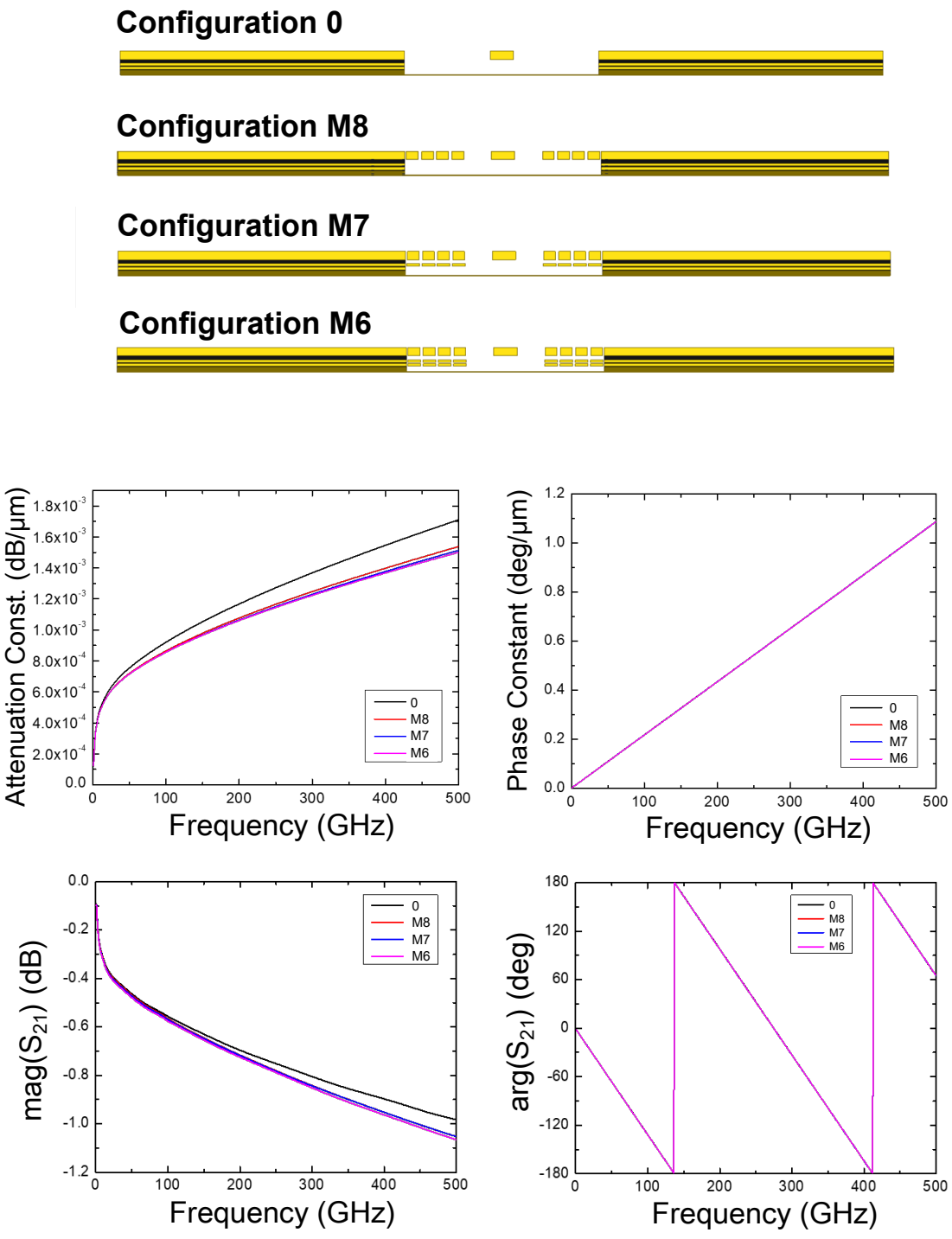


Figure 3.19: L-500G intrinsic simulation performed on different simplified configurations with dummies. "Configuration 0": no dummies (used); "Configuration M8": dummies on M8; "Configuration M7": dummies on M8 and M7; "Configuration M6": dummies on M8, M7 and M6.

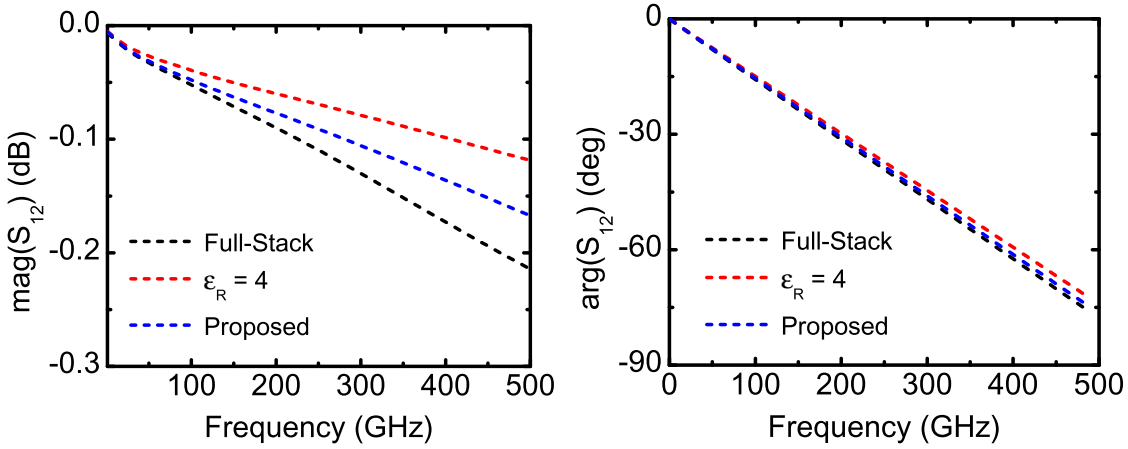


Figure 3.20: Run 2 transmission coefficient intrinsic simulation of thru with different solution for the dielectric stack model. The proposed solution uses three layers: two silicon dioxide with modified properties and a nitride passivation layer on top. Bottom dielectric: $\varepsilon_{R,b} = 3.3$, $\tan\delta_b = 3 \cdot 10^{-3}$; top dielectric: $\varepsilon_{R,t} = 4.3$, $\tan\delta_t = 2 \cdot 10^{-3}$; passivation: $\varepsilon_{R,p} = 4.2$.

acceptable given the reduction in importing and simulation times by several hours.

Since we are interested in the intrinsic element only, we also remove the pads and the access lines of the DUT at this stage. However, the simplified contact layer between M1 and the silicon substrate has been kept in all model configurations.

At this point, the metal and via layers of the simplified element are imported into HFSS. During this step, the heights of each and every layer are assigned as provided by the foundry: the 3D model is generated. The GDSII file contains the 2D dimensions needed to impress the photo-resist: the 3D model needs to be scaled by a ratio of 1.1 in the x and y direction, considering the layout shrink by 10%. Next, we define copper as the material for metals and vias (although each fabricated metal layer has indeed its specific well-defined conductivity) and we include the dielectric and the silicon bulk below the contact layer and a vacuum (or air) box containing each part of the model. We pay attention to the definition of the dielectric layers where the copper stack is immersed.

As said, because of the number of layers with different material properties (30 layers are present in run 2), we decided not to import them all (risking to increase the simulation time, particularly for model configurations with multiple structures). Instead, the layers should be reduced to fewer blocks with modified properties. The easy way is to replace the stack with a single block of silicon dioxide ($\varepsilon_R = 4$): the simulated effective dielectric constant $\varepsilon_{R,\text{eff}}$ stabilizes around 3.6 with similar frequency dependence to the "full-stack" case. However, when looking to the trend of S_{21}, we find that they diverge (see Fig. 3.20). In order to get a better copy of the full-stack transmission coefficient over frequency, while making the model simpler, we decide to optimize the value of $\varepsilon_{R,\text{eff}}$ and the loss tangent $\tan\delta$ (the parameter linked to the losses within the dielectric). After several tests of combinations we were able to find the one proposed in Fig. 3.20, that improves S_{21} and conserves a similar value of $\varepsilon_{R,\text{eff}}$. The couple of permittivity values has been found via an harmonic average on all the dielectrics' heights and relative permittivities of the stack. The deviation from reference is 0.05 dB at 500 GHz, which is well below the measuring instrument sensibility.

Going back to the intrinsic model, now we just need to add the radiation boundary and the excitation ports. A radiation boundary is assigned to the air-box allowing waves to radiate infinitely far into the space. The wave ports for modal solutions are assigned on a planar face tangent to the central connector where the reference plane is located, with an integration line directed from the central conductor to the ground plane (below M1), like in classic microstrip simulations. The dimensions of the ports are such as to cover a portion of space big enough for the E-field to not decay significantly within them: conservatively, our ports extend to the side grounds. The model is now ready to be simulated and results exported (Fig. 3.21).

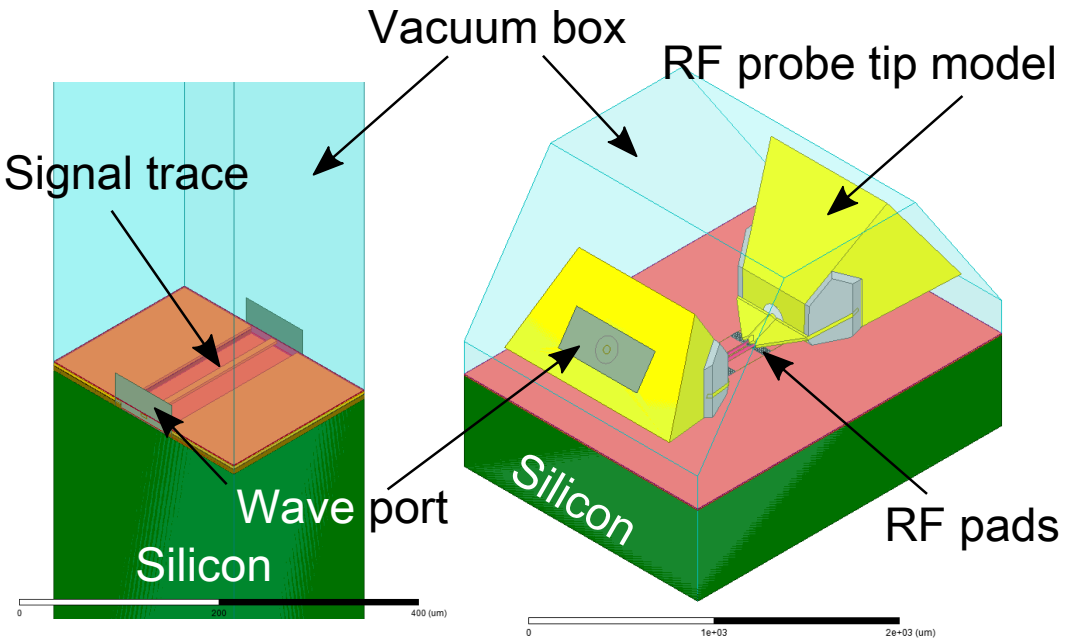


Figure 3.21: Intrinsic (left) and complete probe (right) models for EM simulation of the L-500G (run 2).

3.2.2 Probe Electromagnetic Simulation

Although the intrinsic simulation can be a reliable tool for investigating the measurement at a first glance, it may show severe discrepancies with respect to the actual calibrated data. The origin of this lies on the complex interactions between the measuring instrumentation and the DUT. To understand this behavior, it is necessary to include a RF probe tip model into the EM simulation environment. In other words, the imported circuits are excited with a partial probe model in a much similar way to that of a real-world measurement setup. For instance, a similar approach has been taken by Muller *et al.* [70] that realized a probe model of IP-220 and used it to evaluate the probe tip coupling to substrate on a test GaAs transmission line and on another passive monolithic millimeter-wave integrated circuit (MMIC) and concluded that part of the differences between measurement and intrinsic simulation were due to a resonant mode below the RF pads: only a simulation of this kind could highlight this type of setup influence. Thanks to pictures taken by light microscope, several GSG models replicating the GGB Picoprobes RF probes with a 50/100 µm pitch for each band up to 500 GHz have been designed in HFSS [144] (Fig. 3.22). The probe model is added to the 3D model once the metal stack (pads and access lines included) is imported to HFSS. The shape of the excitation ports and vacuum box are changed according to the new environment, yet the procedure (from creation of the model to simulation) is quite the same (Fig. 3.21).

Once the DUT simulated, however, the exported data are unexploitable since they include the spurious contribution of probes and need to be calibrated. The same procedure of importing and simulating the complete structure with probes has to be repeated for each calibration standard involved into the chosen calibration algorithm. Once all the simulated data collected, they are imported into our IC-CAP calibration "toolkit". Much like measurements, error terms are found thanks to the simulated raw data of each standard and applied to the raw data of the DUT (a verification test structure, for instance). The resulting calibrated S-parameters will be comparable to measurement, since they also embed the electro-magnetic interactions between the DUTs and the environment attributable to the probes, allowing to identify unexpected measurement behaviours not found in the intrinsic simulation.

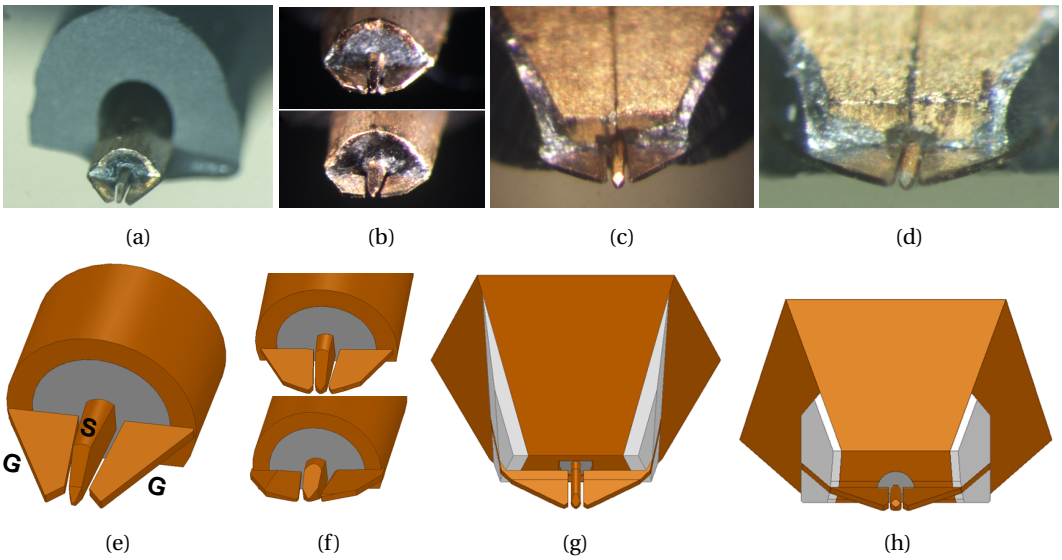


Figure 3.22: Collection of probes for measurement in 1-110 GHz ((a) picture; (e) model), 140-220 GHz ((b) picture; (f) model –port 1 probe has longer pins than port 2), 220-330 GHz ((c) picture; (g) model), and 325-500 GHz bands ((d) picture; (h) model). All probes are GGB Picoprobes® RF probes with G-S-G configuration (shown in (e)). The pitch is 50 µm, except for 1-110 GHz probes (100 µm).

3.2.3 Device Model + Probe Co-Simulation

For the active devices, using a compact device model simulation results into an accurate description of the intrinsic transistor behavior. Compact models of transistors (such as BSIM and PSP for MOS transistors or others more oriented to bipolar transistors, like MEXTRAM and the one used by our team, HICUM, just to mention some) have been added to SPICE simulators to predict the behavior of a circuit design.

HICUM (High-Current Model [103], shown in Fig. 3.23) is optimized for circuits using Si, SiGe or III-V based processes and is particularly accurate at high-frequencies and high-current densities and for this reason it includes a precise description of charges as well as capacitances and transit times as a function of bias. Indeed, this version of HICUM (Level 2) addresses high current, non quasi-static effects, self-heating and avalanche breakdown. However, the model does not cover the entire frequency range up to 500 GHz, as the accuracy above 110 GHz is not guaranteed. Here, measurements up to 500 GHz can be used to fine-tune the transistor's parameters. A verification of the model performance through DC and RF measurements is made in Fig. 3.24.

The presented simulation, however, does not take into account the measuring environment (consisting of probes, pads, BEOL, etc...), just as in the case of intrinsic simulation, and active structures cannot be imported to HFSS, since it does not treat semiconductor equations. Hence, neither way we are able to understand whether the observed measured curves present artifacts generated by the probes or rather due to other causes; most importantly, since the model is not adapted for above 110 GHz, one would be tempted, based on measurement results to fine-tune the HICUM equations to match any unforeseen trend based by measurement.

For instance, the non quasi-static parameters of the transistor (e.g. the ALIT parameter, modelling the delay between the intrinsic base-emitter voltage and the current source, or ALQF, which models the vertical NQS effect on the diffusion charge, or the fcrbi parameter, which is required for lateral NQS modelling or substrate-related parameters [100]) cannot be extracted at LF, and may give the best fit if and only if a complete sub-millimeter measurement is performed, after double-checking by some sort of complete probe simulation that any unexpected trend is indeed replicated. For this purpose, an hybrid solution has been imagined [34] by which, for the first time at the best of the author's knowledge, we are able to evaluate the effect of probes on the measurement accuracy of the FoMs of a transistor.

This solution embeds the compact model to the HFSS environment; the complete-open HFSS

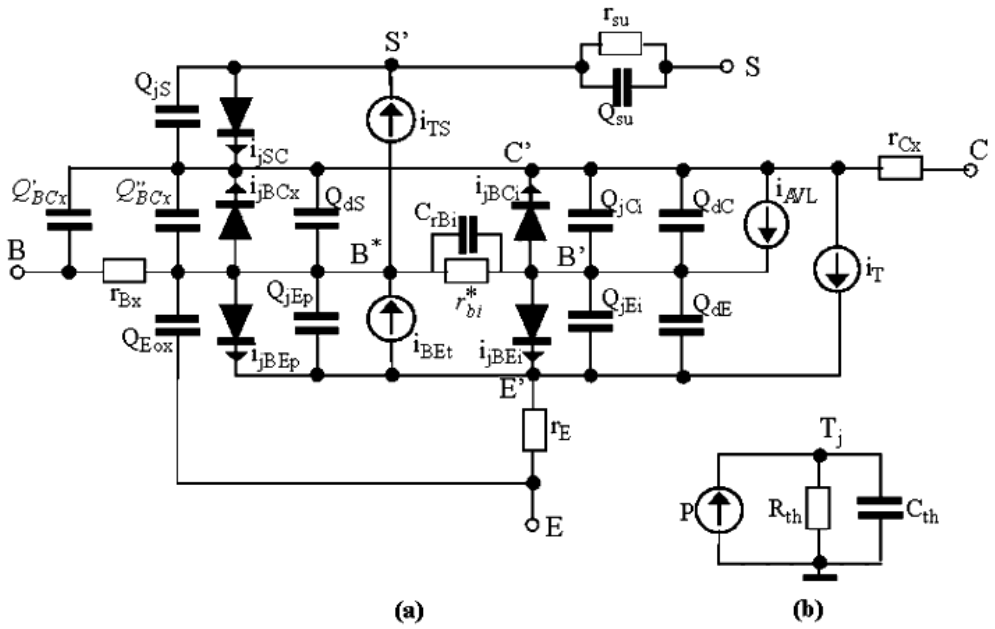


Figure 3.23: HICUM model – equivalent transistor circuit (courtesy of [115]).

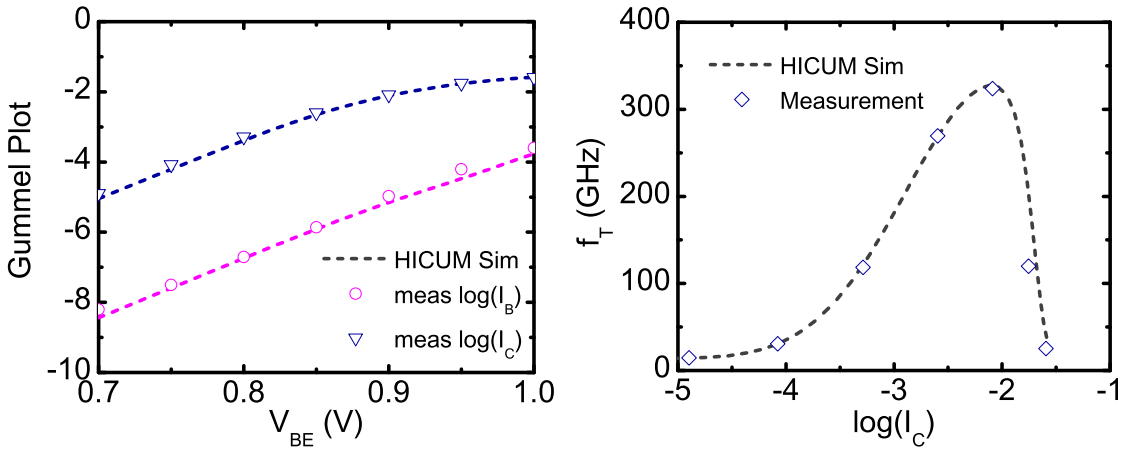


Figure 3.24: Comparing the HICUM model to measured data (run 2) as a function of DC quantities: gummel plot and transit frequency vs. collector current (extracted at 40 GHz). $V_{CB} = 0\,V$.

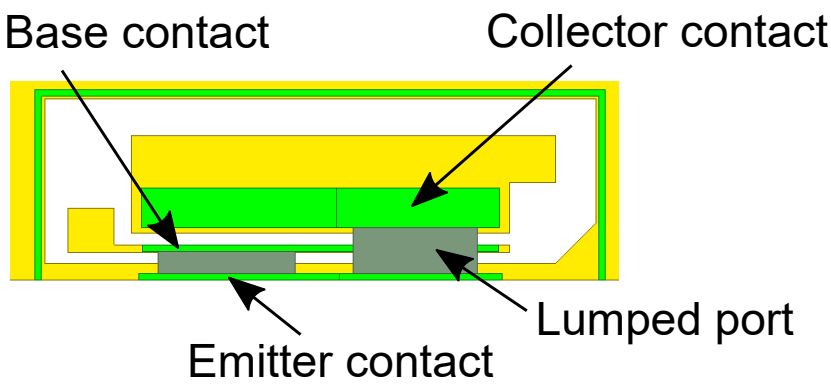


Figure 3.25: Connections for EM+HICUM co-simulation: lumped ports beneath contacts. Color key: M1 in yellow, contact pins in light green tone, ports in green. Only half of the transistor's layout is shown.

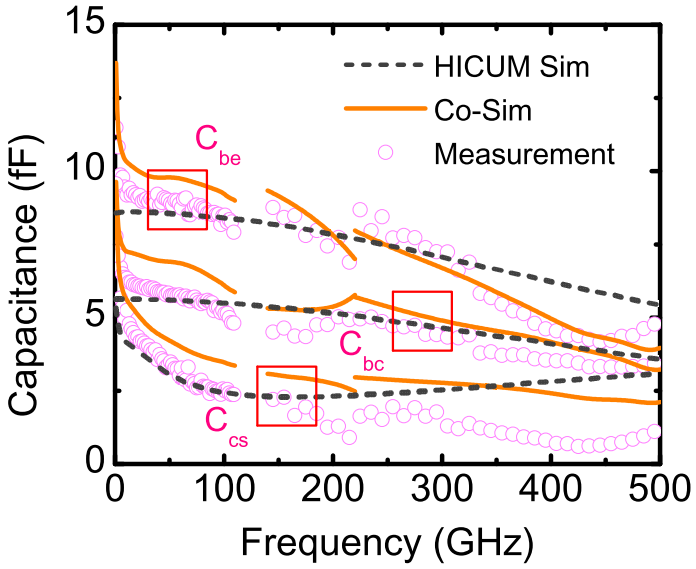


Figure 3.26: Transistor capacitance analysis comparing the HICUM model to both co-simulation and measured data (run 2). $V_{CB} = V_{BE} = 0\,V$.

model has been modified and two extra lumped ports have been added beneath the contact pins which are normally absent in its layout (see again Fig. 3.14). This is made to output the EM signal at the position where the base and the collector should be located (Fig. 3.25), bringing the HFSS model to a total of 4 ports (two probes wave ports + base lumped port + collector lumped port). The resulting 16-terms matrix representing the S-parameters of the measuring environment is then linked to the compact HICUM model, through the well-defined base and collector ports. The joint electro-magnetic + HICUM simulation, named *co-simulation*, is performed in IC-CAP after that the transistor circuit model has been simulated with proper parameters (experimentally validated values of capacitance, resistance, transit time, parameters accounting for self-heating and non-quasi static effects, etc...). To take into account the use of different probe sets in the real world, we perform it for each probe model: 4 sets of data are therefore produced for a single DUT, spanning from 1 to 500 GHz. These data are subsequently assembled in IC-CAP and calibrated with the complete models' simulated data of each calibration standard, much like for a verification (passive) DUT.

Preliminary, we show here the co-simulation applied to the cold measurements of the HBT: $V_{CB} = V_{BE} = 0\,V$, transistor-off (Fig. 3.26). Simulated values are akin and measured trends deviating from the HICUM simulation alone (all the bends in the first range and the HF bend of C_{be}, as well as trends in the 140-220 GHz band) are largely explained by the presence of probes and subsequent calibration. C_{cs} HF measurement deviation will be better explained by the next observations on other measured parameters.

3.3 Calibration Toolkit

Thanks to the presented structures, we are effectively able to perform SOLT and TRL calibration on the test devices. As already mentioned, the full description of the TRL calibration algorithm (the one mainly used in this work due to its effectiveness at HF) is given in Appendix B. This algorithm has been implemented by our team in Keysight's Integrated Circuit Characterization and Analysis Program (IC-CAP). The code takes as inputs the raw measured (or simulated) data of the calibration standards as well as raw data of the DUTs (either the HBTs themselves or any other structure not involved in the calibration process) and automatically outputs the calibrated (and, if necessary, de-embedded) results, in a single unified set of data, for multiple frequency bands and bias (when present). It is possible to adapt the solutions to the calibration technique

employed, set of probes, geometry of lines and desired position of the reference plane, and it is possible to choose between any of the measurement campaigns or set of measurements one would wish to investigate, in a specific or in the whole frequency range. In particular, our toolkit is able to perform an additional manipulation to the data: the characteristic impedance correction. In this section, we will discuss one of the most critical points of the TRL calibration, i.e. its reference impedance, which is equal to the complex, frequency-dependent characteristic impedance of the fabricated transmission line.

3.3.1 Characteristic Impedance Correction

After having performed the TRL calibration on a given raw set of data $[T_M]$, with raw data of thru, lines (L-110G and L-500G) and reflect (P-O or P-S), we compute the error T-matrices $[T_X]$ and $[T_Y]$, and finally with the definition of the position of the reference plane, through the $[T_h]$ matrix, we obtain the "intrinsic" T-matrix of the DUT $[T_{DUT}]$, its T-parameters as seen from the reference plane:

$$[T_{DUT}] = \left[[T_X] \cdot [T_h]^{-1}\right]^{-1} \cdot [T_M] \cdot \left[[T_h]^{-1} \cdot [T_Y]\right]^{-1} \tag{3.2}$$

At this point, however, we have still not taken care of the characteristic impedance. If, as it is inevitably the case, the characteristic impedance of our designed line at the port defined by the reference plane does not match the system impedance, this could result in reflections of the injected/received signal, thus leading to measurement quality loss.

Research studies have explored this topic. Eisenstadt [26] proposed an analytical method for the extraction of γ and Z_0 starting from the ABCD matrix of a general lossy unmatched (intrinsic) transmission line and the hypothesis of a controlled microwave system (with reference impedance $Z_0 = 50\ \Omega$). He could come to two straightforward formulas for γ and Z_0 which depend on the S-parameters of the line only. This method can provide an approximation to the values of γ and Z_0 (and the R, L, C, G parameters) of the line, since it does not account for the measurement test fixtures, nor probe non-idealities or electrical transitions. This method can provide, indeed, only a fair estimation of the line electric parameters.

Another method, based on the comparison between two calibration techniques was explored by Williams, Marks *et al.* [127, 63] and later expanded by the same authors [128]. It consists of using data from two measurement sets where the reference plane is the same for the two calibrations and one has a well-known characteristic impedance: if the second calibration is the (m)TRL calibration, the calibration reference impedance will coincide with the characteristic impedance of the line. This method is particularly well-suited for lines printed on silicon and other lossy substrates and is insensitive to even large shunt pad admittances [129, 130]. However, both the previous methods demand a direct comparison to "easily characterized" reference lines, i.e. located on a calibration substrate (off-wafer); e.g. in [129] they are located on an ISS and are calibrated with the lumped-load method (the one we are introducing next and, in fact, the one we opted for). Moreover, we employ a low-loss substrate ($\varepsilon_{r,eff} \approx 3.5$) and an homogeneously designed line: the calibration comparison would be unnecessary cumbersome in the case of our study.

Galatro and Spirito [39] proposed an original method for the characteristic impedance extraction, feeding the EM simulation-based value of Z_0 into both the previous algorithms, avoiding using any off-wafer equipment. However, even though this "*a priori*" extraction performed well also with respect to pad-to-line discontinuities, it necessitates of additional simulation setup and it is less straightforward than the approach used here.

The methodology employed in our work is explained in the following. The pursued approach was not to use complex impedance extraction routines nor synthetic (i.e. simulated) data, but to exploit on-wafer measurement with straightforward data manipulation; a full comparison with such different approaches may be tackled in future. As pointed out by Marks and Williams [65], the calculation of the characteristic impedance of the line can result from knowing the propagation constant γ and an estimate of the line capacitance. In fact, as explained in Appendix B, γ is found

as a by-product of the TRL calibration algorithm from [27]. Its value depends only on the difference between the lengths of the line and the thru, $l_L - l_T$, and on the product of the T-matrices of the line and the thru.

From the theory of transmission lines, we know that:

$$\gamma = \alpha + j\beta = \sqrt{(R + j\omega L)(G + j\omega C)} \tag{3.3}$$

where α is the attenuation constant and β the phase constant, and R, C, G, L are the lumped electrical elements of the line. Moreover, the characteristic impedance of a general lossy line, from theory again, can be described by:

$$Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \tag{3.4}$$

By taking the ratio of Eq. 3.3 and 3.4, we can write:

$$\frac{\gamma}{Z_0} = G + j\omega C = j\omega C\left(1 - j\frac{G}{\omega C}\right) \tag{3.5}$$

Note that C and G are unknown. If we can determine these quantities, then the characteristics impedance can be calculated from the relations. Two hypothesis are made to simplify this relation:

- C should be nearly independent of frequency and metal conductivity;
- $G \ll \omega C$, which means that the losses in the dielectric substrate of the line should be negligible compared to the reciprocal of the reactance of the line.

At this point, we consider again Eq. 3.5 and with the hypothesis holding true, we simply find an estimation of the characteristic impedance Z_0 by:

$$Z_0 \approx \frac{\gamma}{j\omega C} \tag{3.6}$$

In order to obtain the only remaining parameter of this expression, namely the DC line capacitance C, some solutions have been proposed [65, 134], and we designed a load standard to pursuit the so-called "lumped-load method" introduced by Williams and Marks [134]. The raw measurement (or simulation) of the pad-load are used for the sake of properly matching the impedance of the system to that of the lines. To use this approach, we establish a third assumption in addition to the previous two:

- $Z_L \approx R_{L,dc}$, i.e. the impedance of the load should be approximately equal to its real part at DC. This condition is valid for small lumped resistors, like the one implemented in the pad-load, and the main contribution to the load reactance comes from the equivalent reactance of the via stack [93].

So if we consider the expression of the load capacitance, by applying this last condition:

$$Z_L = Z_0 \frac{1 + \Gamma_L}{1 - \Gamma_L} \approx R_{L,dc} \tag{3.7}$$

where Γ_L is the reflection coefficient of the load, computable once the error terms of the TRL calibration have been found by the algorithm. Eventually, by rearranging Eq. 3.5:

$$C\left(1 - j\frac{G}{\omega C}\right) = \frac{\gamma}{j\omega}\frac{1}{Z_0} \tag{3.8}$$

and finally combining with 3.7, we obtain, with the previous hypothesises holding true:

$$C \approx \mathrm{Re}\left(\frac{\gamma}{j\omega}\frac{1}{R_{L,dc}}\frac{1 + \Gamma_L}{1 - \Gamma_L}\right) \tag{3.9}$$

At this point, we put this value into Eq. 3.6 to calculate the characteristic impedance, and we are now capable to consider the transformation from this characteristic impedance at the reference plane to the desired 50 Ω impedance as a S-matrix for line with unequal terminating impedances, i.e.:

$$[S_z^{+/-}] = \frac{1}{Z_0 + 50} \begin{bmatrix} \pm(50 - Z_0) & 2\sqrt{50\,Z_0} \\ 2\sqrt{50\,Z_0} & \pm(Z_0 - 50) \end{bmatrix} \tag{3.10}$$

Finally, Eq. 3.2 has to be adapted, when the impedance correction is applied, and becomes:

$$[T_{DUT}] = \left[T_z^+\right]^{-1} \cdot \left[[T_X] \cdot [T_h]^{-1}\right]^{-1} \cdot [T_M] \cdot \left[[T_h]^{-1} \cdot [T_Y]\right]^{-1} \cdot \left[T_z^-\right]^{-1} \tag{3.11}$$

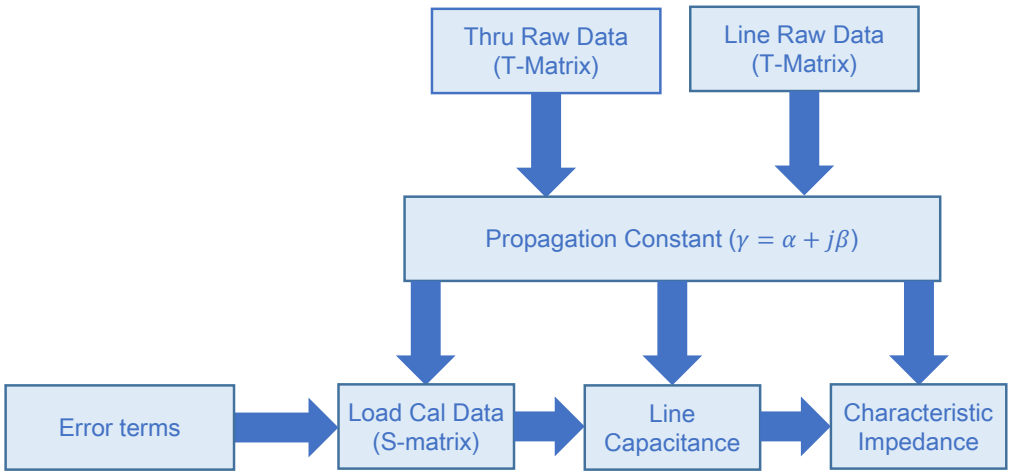


Figure 3.27: Flowchart of the extraction of the characteristic impedance for its correction.

The three conditions upon which the lumped-load method is based are best verified at LF because essentially they assume the line capacitance equal to its DC value and low losses in the substrate. These hypothesis are hard to rigorously validate at HF and may result into an inappropriate correction. We investigate now what can be the effects of an impedance correction with different degrees of approximation based on the lumped-load method, in order to find what should be the approach to use next.

In Fig. 3.27 we sum up how our toolkit retrieves all the parameters of the line. We can see that the extraction of the propagation constant γ is straightforwardly made by a matrix multiplication of two sets of raw data (from Eq. B.11 and B.16), namely the raw data of the thru and the line, once the S-parameters have been converted to T-parameters (chain transfer matrix).

In Fig. 3.28 we present α and β computed by using data from both the lines: theoretically they should provide identical results since they share the same (traversal) geometrical and material properties. These curves are juxtaposed to the curve from the intrinsic simulation of the line, for verification. Indeed, we can observe that the propagation constant follows the linear trend of the intrinsic trace and at high frequency just small fluctuations appear around the intrinsic value.

The attenuation constant turns out to be harder to evaluate at HF, with a non-physical shape from 400 GHz. A comparison of the same configuration with neighbors has been made and no impact from the adjacent structures has been observed. Also, probe contact repetitiveness can be excluded, since the same behavior has been noticed in different measurements. We have found, however, that L-500G, if calibrated, showed a hump around the same frequency range (400 GHz on) on mag(S_{11}) and simultaneously mag(S_{21}) > 0 dB. For helping in drawing conclusions on the propagation constant, let us refer to Fig. 3.29, where probe simulations are juxtaposed to the previous measurements. Although not fully replicated, the shapes of each curve's trend are quite

similar: for instance, on α, see the rather flat trend in the first band, the rise in the second and the dip in the last band, as well as the slight divergence on β, second band. However, the magnitude of the measured curves keeps distant from both simulated curves, indicating some additional effects unaccounted for but hard to be addressed by such a partial analysis (e.g. contact inaccuracy, crosstalk, etc...). The calibration will be eventually affected by this incorrect performance of the line at HF. We will conclude and try to attribute a cause to that later in this manuscript.

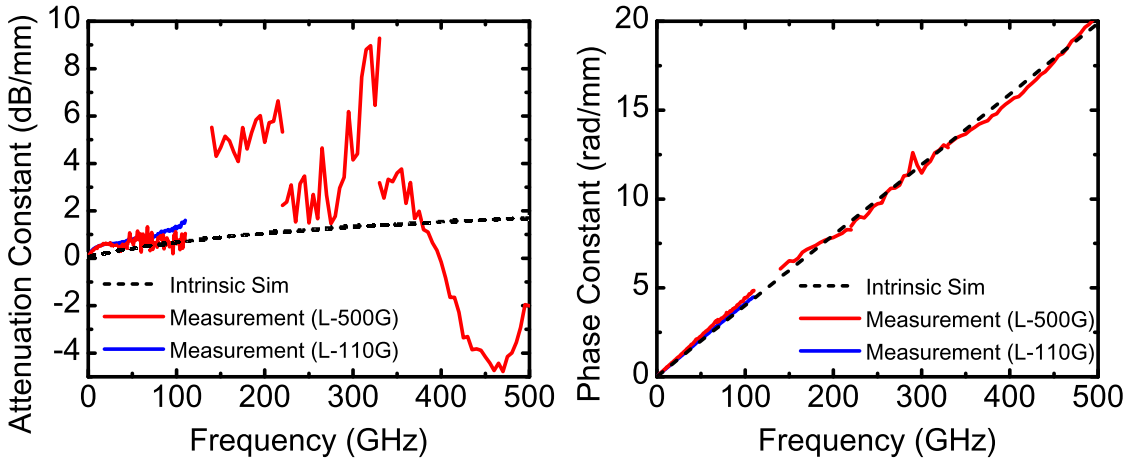


Figure 3.28: Run 2 attenuation constant α = Re(γ) and phase constant β = Im(γ) found for L-500G and L-110G: intrinsic simulation (dashed) and measurement from point-by-point data (solid). Data of L-110G have not been measured above 110 GHz.

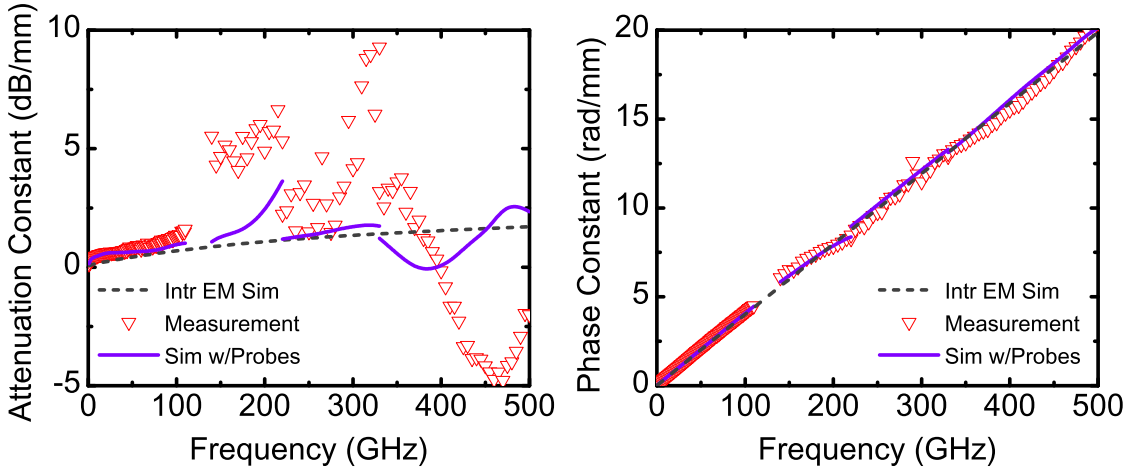


Figure 3.29: Run 2 attenuation constant α = Re(γ) and phase constant β = Im(γ) found by measurement and complete-probe simulation. Measurement are from point-by-point data.

From Fig. 3.27, it is clear that in the calculation of the line capacitance, and from that the characteristic impedance, the complete TRL error term matrices $[T_X]$ and $[T_Y]$ need to be known, in order to calibrate the load standard (pad-load) and be able to use its reflection coefficients Γ_L in Eq. 3.9. The line capacitance calculated in this way is shown in Fig. 3.30, and compared to the intrinsic simulation, the extracted value just exceeds (at its worst) 15% of it. We can also check that the adimensional ratio $G/\omega C$ keeps below 0.01 and yields 0.003 at LF (below 1 GHz), thus verifying one of the conditions for applying the lumped-load method.

Because of the multitude of mathematical operations involved to find Z_0, one may be tempted to simplify the algorithm in order to find an approximate matrix $[T_z]$ to feed into Eq. 3.11. Three leads are followed.

- The first one (see the green line in Fig. 3.31) takes a constant value of the line capacitance

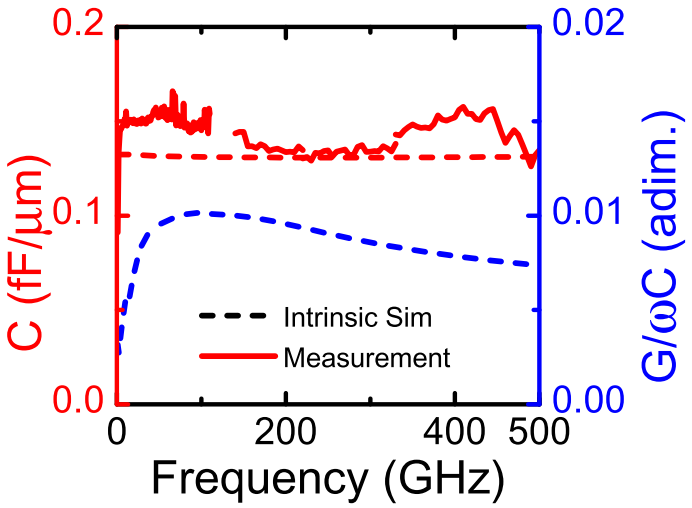


Figure 3.30: Run 2 line capacitance and verification of the constant C and $G/\omega C \ll 1$ conditions: intrinsic simulation (dashed), measurement from point-by-point data (solid red).

in a limited frequency range, namely a low-frequency portion of the spectrum where we are quite confident that HF effects do not come into play (below 110 GHz). In the case of these specific measurements, the discrepancy at LF of the line capacitance from the intrinsic curve (Fig. 3.30) suggests an inaccurate measurement, although this is not a common situation. We expect, consequently, the correction in such a circumstance to be partially inadequate. The very first frequency points of the first band are also excluded, since they are very noisy and they are outside of the range of validity set by the lines.
We therefore consider for the computation of C the range from 27 to 55 GHz, and we take a mean value of the slightly frequency-dependent C: in the example (run 2) we have $C_1 = 0.1395\,\mathrm{fF/\mu m}$ for port 1, $C_2 = 0.1398\,\mathrm{fF/\mu m}$ for port 2. By holding the line capacitance constant, the value of Z_0 over frequency is only dependent from γ (from Eq. 3.6).

- The second one (blue line, Fig. 3.31) derives from the observation that the phase constant β, as we can see in Fig. 3.28, and as it is also defined in Eq. 3.1, is linearly dependent with frequency, and can be expressed by $|\beta| = \frac{\omega\,\tau_\phi}{l_L}$. By rearranging Eq. 3.6, and neglecting alpha, we obtain:

$$Z_0 \approx \mathrm{Re}\left(\frac{\alpha + j\frac{\omega\,\tau_\phi}{l_L}}{j\omega C}\right) = \frac{\tau_\phi}{l_L C} \tag{3.12}$$

and we are basically removing the attenuation losses on the line by just considering the real part of Z_0, in this very rough correction approach. We can therefore imagine to take just an extracted constant value for τ_ϕ and feed it, alongside the constant C value, in the formula for the Z_0 extraction (Eq. 3.12). By a simple interpolation of β in the same frequency range of extraction of C, we get $\frac{\tau_\phi}{l_L} = 6.7\,\frac{\mathrm{ns}}{\mathrm{m}}$, which yields $46.69\,\Omega$.
- the third approach cannot be visualized in Fig. 3.31, where we show only the real part of Z_0. In fact, the previous approach is developed by assuming that the line is lossless. As we see in Fig. 3.28, however, this is not the case, and losses are indeed present, even though, as displayed by the actual intrinsic line, they are likely to stay low since α, after deduction of its unstable calibrated/measured trend, grows to no more than 2 dB/mm at 500 GHz and at most to 4 dB/mm according to the complete probe simulation.
Even so, α, which is related to both losses on copper and the dielectric, does need to be considered for impedance correction matrix calculation, since the impact of the imaginary part of Z_0 won't be negligible. $\mathrm{Im}(Z_0)$ will therefore be included into the $[T_z]$ correction matrix.

$Re(Z_0)$ is retrieved as described in the second approach.

The results are presented in Fig. 3.31: the first two methods are compared alongside the intrinsic simulation (dashed curve) and the extracted value from the complete algorithm which takes the frequency-by-frequency (point-by-point) values of both C and the phase constant β (red solid curve). We observe that the "constant C" method provides a generally less noisy and more continuous curve (in particular in the first band) due to the fact that the additional matrix calculations to find C at every point are avoided. The constant value resulting from the "constant C, τ" method is 1.5 Ω below the target intrinsic at HF: 48.3 Ω, and inevitably falls short in replicating the trend in the lower part of the spectrum.

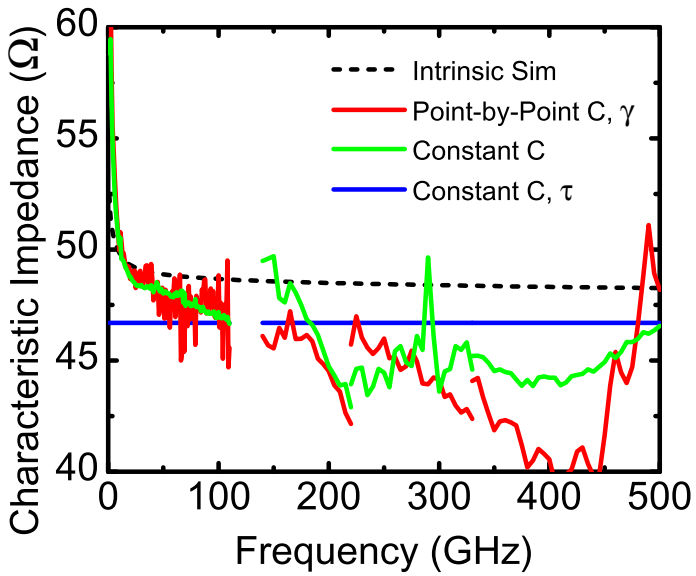


Figure 3.31: Run 2 line extracted real part of the characteristic impedance: intrinsic simulation (dashed), measurement from point-by-point data (solid red), measurement from a constant C value (solid green), and measurement from constant C and τ_ϕ values (solid blue).

Nevertheless, these conclusions on the Z_0 correction do not tell irrevocably which method would provide the best results on an actual DUT. We are going to compare the previous methods for the Z_0 extraction on the measurements of one of the introduced verification devices, namely run 2 complete-open (no-T0). We will extract the port capacitances of this device, that are indicative of the trends of all the measured quantities since they are a combination of all the S-parameters. The results are presented in Fig. 3.32 for port 1 capacitance C_{11} and port 2 capacitance C_{22}. The figures also put in comparison the case where no correction at all is applied to the characteristic impedance. We present the relative deviations in percentage bench-marking the reference value (intrinsic), to improve the visibility of the plot.

While in the first band (up to 110 GHz) the actual "point-by-point" extraction of Z_0 (complete algorithm) provides better results by around 5%, in all the other bands the "constant C and τ" (considering an ideal lossless line, $\alpha = 0$) outperforms the other two and the error even settles within 5% in the 240-480 GHz range. However, it can be stated that avoiding correction does not expose our measurements to unrealistic values, inasmuch as its trend is comparable to the others (even 2% better, on average, than the best correction in the 20-110 GHz band). Considering $Im(Z_0)$ ($\alpha \neq 0$) the previous conclusions do not change, as we can observe in Fig. 3.33, since the orange line is perfectly superimposed in both capacitances to the blue line.

To notice the importance of correcting the impedance by including the ohmic losses, we need to turn our focus toward more complex parameters. Fig. 3.34 shows the usual two main figures of merit of an HBT, f_T and f_{max}. Both the co-simulated and measured data are calibrated with TRL and de-embedded with C-S and C-O. Moreover, the data are impedance-corrected by the "constant C,

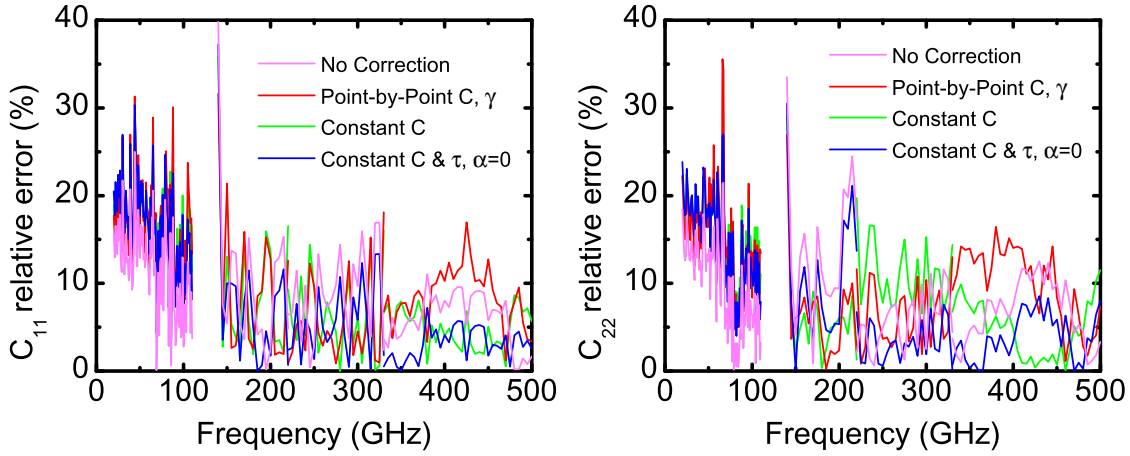


Figure 3.32: Run 2 complete-open (no-T0) port 1 and 2 measured capacitance's relative deviation, referred to the reference intrinsic simulation. Uncorrected data compared to corrected data with several methods for Z_0 extraction: "point-by-point" extraction, "constant C" method, "constant C and τ_ϕ" method.

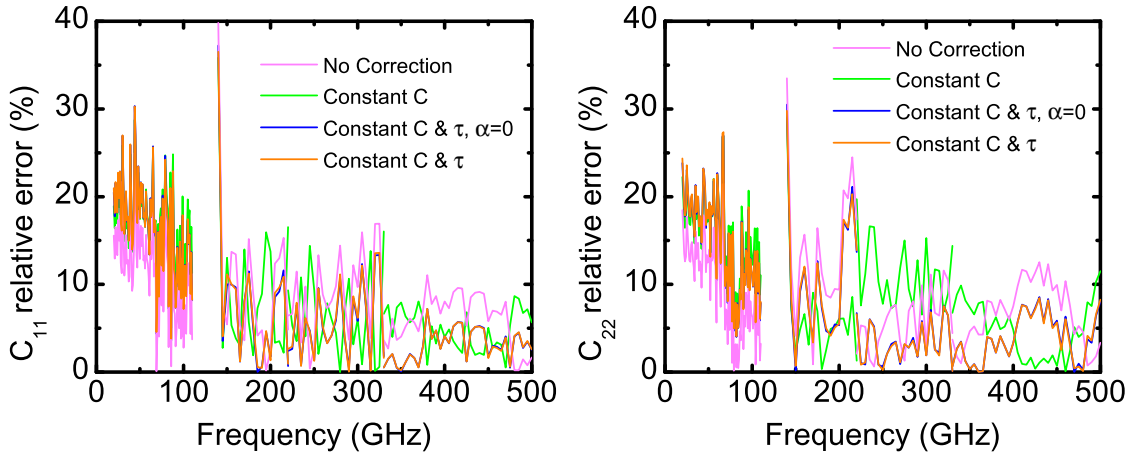


Figure 3.33: Run 2 complete-open (no-T0) port 1 and 2 measured capacitance's relative deviation, referred to the reference intrinsic simulation. Uncorrected data compared to corrected data with several methods for Z_0 extraction: "constant C" method, "constant C and τ_ϕ" method and "constant C and τ_ϕ" including α.

τ" method, but in one case with just the real part of Z_0 ($\alpha = 0$), whilst in the other also including its imaginary part.

Again, curves are perfectly superimposed on f_T. However, because of the dependence of f_{max} to the base series resistance (and therefore that of the access line expressed by α), the impedance correction which includes α yields a completely different curve, closer to the HICUM model, in particular at LF. This correction is eventually important even when nominally 50 Ω lines are used, since the actual characteristic impedance is large as the surface resistance is high at LF, hence the resistance per unit length [132, 84].

In conclusion, we have seen that the transistor measurement (via f_{max}) are particularly sensitive to an impedance changing, whereas for our verification standards, short and open, the alpha correction is less important, possibly because, in these structures, the wave will always be reflected, and we cannot appreciate the effect of matching.

Thanks to its easy computation and limited data manipulation, and because of its completeness, the "constant C, τ" method including α proves to be the best candidate for Z_0 extraction in the whole spectrum, and will be used hereinafter to correct the measurements.

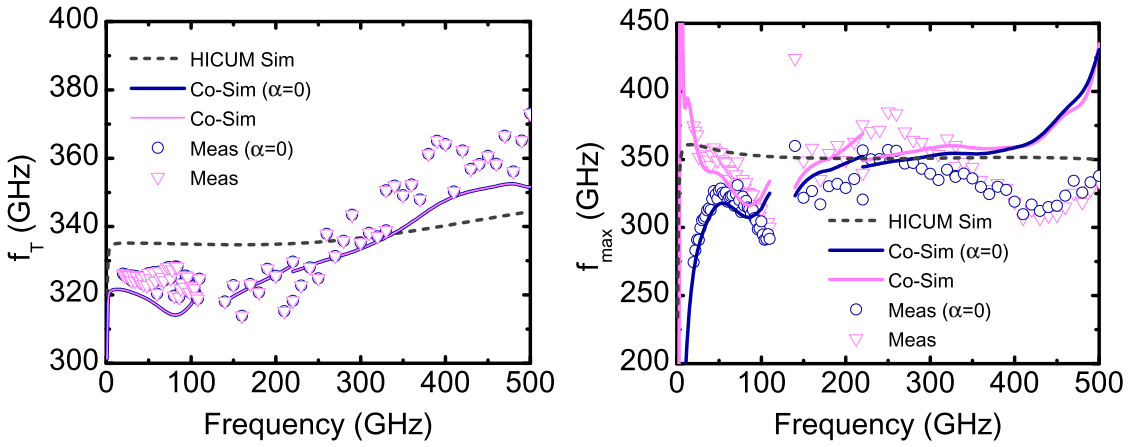


Figure 3.34: Transit frequency and maximum oscillation frequency (run 2). HICUM simulation compared to both measurement and co-simulation data of the HBT, calculated by applying two methods for Z_0 extraction: "constant C and τ_ϕ" method and "constant C and τ_ϕ" including α. $V_{CB} = 0\,V$, $V_{BE} = 0.9\,V$.

3.4 SOLT vs. TRL Calibration Approach

Once all the structures and analysis methods introduced, we now want to use our on-wafer calibration toolkit by comparing different calibration methods to the intrinsic simulation of a verification standard chosen among the designed structures on the die. The calibration algorithms we will consider are SOLT and TRL, both implemented in our toolkit.

Similar HF comparisons have been made in the past; however, the following differs in several ways. Williams *et al.* [132] provided an exhaustive comparison of different calibration and de-embedding methods (TRL vs. SOLT vs. LRRM with open-short or thru-line-short-open de-embedding), however their analysis was limited to 110 GHz. Also, it should be remarked that the load used by the authors for the impedance correction is a trimmed 50 Ω load from a calibration substrate, while in this work we implemented it directly on-wafer. Fregonese *et al.* [32] performed a more frequency-extended comparison (up to 500 GHz) between TRL and ISS SOLT (and ISS TRL) calibrations, employing EM simulations to validate the calibration substrate extracted values and the measurements themselves. However, simulations relied on a simple "ideal" coplanar probe and measurements and analysis stem from a different technology. The simulation apparatus has been improved since then with dedicated probe set models and the measured results we present here come from the novel run 2 layout. To evaluate the performance of an on-wafer calibration compared to another made on a different host medium, like a thin fused silica substrate, we will perform two kinds of SOLT calibrations: ISS SOLT, oftentimes incorrectly used to refer to a SOLT performed using a general calibration substrate, and, for the first time, on-wafer SOLT up to 500 GHz. This former SOLT method is the most widely used in industry for its simplicity and versatility.

Our analysis spans from 1 GHz to 500 GHz; however, due to the loss of accuracy experienced by the ISS SOLT, the results in this case stop at 220 GHz. Table 3.5 presents the substrates reported on the data sheets of each commercial off-wafer substrate, which are necessary to calculate the reflection coefficients of the standards to perform the SOLT algorithm:

- the characteristic impedance of the line, Z_0;
- the capacitance of the open, C_0;
- the inductance of the short, L_0;
- the delay of the thru, τ_ϕ;
- the load inductance, L_{match}, or, alternately, the load capacitance, C_{match}.

These nominal values depend on the type of substrate, but also on design, pitch and configuration of the probes and are defined by the vendors [25]. The commercial substrates we used are, apart from the already presented CS-5 from GGB Industries, used up to 110 GHz with PP-110,

Cal Substrate	Freq Band (GHz)	Z_0 (Ω)	C_0 (fF)	L_0 (pH)	τ_ϕ (ps)	C_{match} (fF)	L_{match} (pH)
CS-5	1-110	50	6.5	5	1.13	3.1	-
138-357	140-220	50	0.7	6.8	0.5	-	9
138-356	220-330	50	5.9	16.5	0.5	-	7.8
Silicon	1-500	50	1	3.5	0.42	-	6.3

Table 3.5: Calibration substrate standards definition with their frequency range of adoption for SOLT calibration.

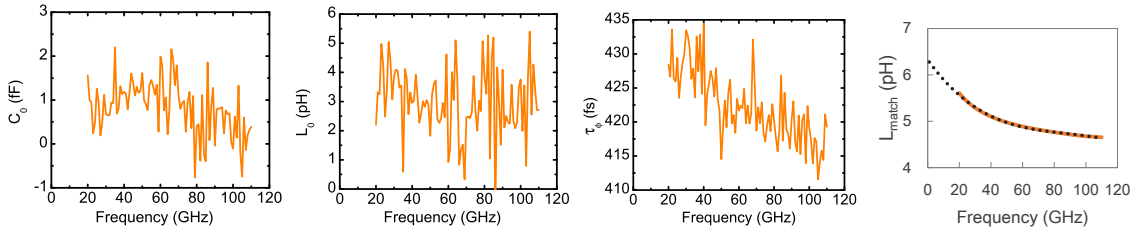


Figure 3.35: TRL-calibrated port capacitance of P-O, port inductance of P-S, and thru delay at the left, P-L inductance provided by intrinsic simulation at the right. Polynomial interpolation (third order) of L_{match} yields a LF value of 6.3 pH.

the 138-357 ISS support from Cascade Microtech, used from 140-220 GHz with IP-220, and the 138-356 ISS support again from Cascade Microtech, used from 220-330 GHz with IP-330.

The on-wafer SOLT calibration has been performed on our run 2 calibration kit, implementing the SOLT algorithm on our toolkit. To define the calibration standards, we used an hybrid simulation-measurement approach. The inductance of the short (pad-short), the capacitance of the open (pad-open), as well as the delay of the thru have been computed from the measured data of these devices, after applying our implementation of the on-wafer TRL calibration (all reported in Fig. 3.35). Our SOLT algorithm, much like VNA's, requires constant (or interpolated) values of the standards' coefficients: since we did not experience any significant improvement with a third-order or lower interpolation, we extracted an average value of them. We tried to follow the same procedure (retrieval from measurements) for the load standard (pad-load) extraction of its reactance, but since the physical loads are located in the silicon substrate below the BEOL, it is difficult to access an accurate value by raw measurement plus calibration. Therefore, we opted for generating the intrinsic load model and retrieve the values directly from an HFSS simulation. Furthermore, we have decided to use only the first term of the third-order polynomial fit for L_{match}, since the higher-order terms did not provide any significant improvement once inserted into our toolkit (interpolation is also shown in Fig. 3.35). The position of the reference plane in each and every of these structures, either measured or calibrated, is the usual position we have employed in our TRL analysis so far (the "b-b" position of the thru, at the edge of P-O).

The following plots show the results found by applying the TRL, ISS SOLT and on-wafer SOLT calibration on different verification structures. The probe sets used are:

- TRL and on-wafer SOLT: PP-110, PP-220, PP-330, PP-500;
- ISS SOLT: PP-110, IP-220, IP-330, PP-500.

Choosing different probe sets should not affect the calibrated results since the calibration algorithms handle precisely those contributions. However, interaction with different materials, different geometries, etc. do have a visible effect on the corrected data. Moreover, depending on the performed data operations, the outcome might be different between the two calibration algorithms, even though the error models upon which they rely are indeed interchangeable. In fact, the cross-talk correction is not performed on a 8-term algorithm such as TRL.

For these methods to be effectively comparable, we have to take care in identifying and possibly correcting the location of the reference plane after calibration. As mentioned, after SOLT calibration, the reference plane is located at the probe tips, while it is positioned at a position on

the thru defined by the user after the TRL calibration. Moving all the reference planes at the probe tips would lead to inaccurate results, since parasitic contributions are higher as more elements are taken into account after the calibration, and the user cannot precisely define the position of the tips on the pads. This is why we performed an additional pad-open/pad-short de-embedding step to the SOLT-calibrated data, that should take into account any L/C due to the pads and access lines. The reference plane is now located at the access line edge, in the same position where we locate it after TRL calibration and also where the intrinsic model has its boundary (position "b" in Fig. 3.3). To sum up, here are the post-measurement steps:

- passive DUT's raw data ⟶ on-wafer TRL calibration + Z_0 correction ⟹ TRL-calibrated passive DUT (ref. plane: edge of access);
- passive DUT's raw data ⟶ on-wafer SOLT calibration ⟶ P-O de-embedding ⟶ P-S de-embedding ⟹ SOLT-calibrated passive DUT (ref. plane: edge of access);
- passive DUT's raw data ⟶ ISS SOLT calibration ⟶ P-O de-embedding ⟶ P-S de-embedding ⟹ SOLT-calibrated passive DUT (ref. plane: edge of access).

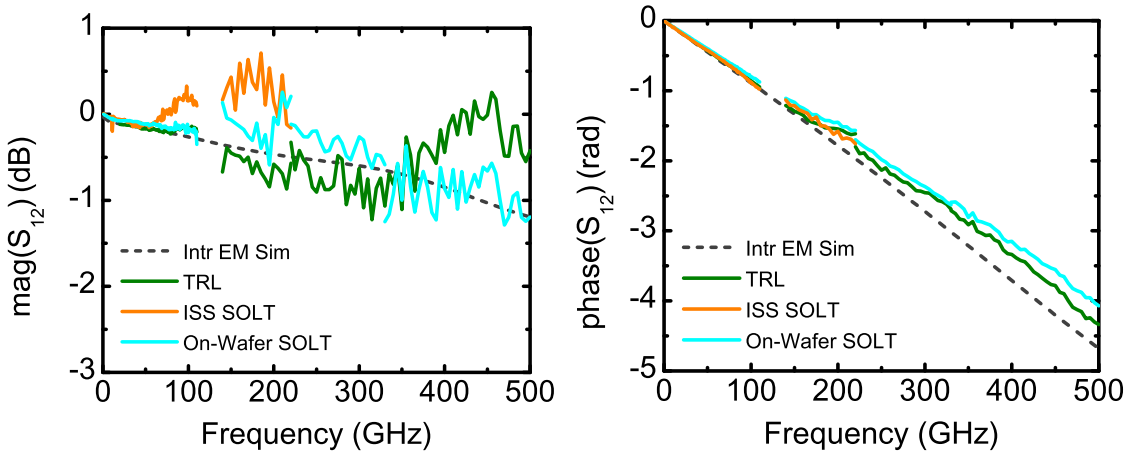


Figure 3.36: Transmission coefficient S_{21} measurements and intrinsic HFSS simulations of L-500G (M) applying different calibration methods (run 2). Note: the reference plane position after TRL calibration (and its intrinsic simulation) and SOLT calibration is the same.

The first verification DUT we consider is L-500G (M). Curves of S_{12} (magnitude and unwrapped phase) are shown in Fig. 3.36. We immediately see the orange curve of the ISS SOLT deviating and taking nonphysical values of magnitude starting from around 60 GHz, and even though the phase better follows the intrinsic line, it deviates considerably in the third frequency range due to the increasing number of different parasitic contributions developed in the two different media. Comparing the two on-wafer methods we find a fairly good replication all over the spectrum, with on-wafer SOLT outperforming the TRL and sticking to the intrinsic trend up to 500 GHz. The TRL calibration however yields the best performance on the phase in the whole considered spectrum region.

The inductances of C-S are now considered (Fig. 3.37). Both ISS and on-wafer SOLT have similar LF trends and are seen underestimating the port inductances compared to the on-wafer TRL calibration in the whole considered frequency range. On port inductances, on-wafer SOLT maintains rather constant, compared to the intrinsic simulation, indicating that the de-embedding on on-wafer SOLT has probably not captured HF parasitics (series inductance). Better results on evaluating the mutual inductance up to 500 GHz are provided, on the other hand, by the on-wafer SOLT, as TRL plots oscillating and almost nonphysical curves, with an imprecise strong offset in the first band.

We suppose L_1 and L_2 to be pretty much symmetrical, as it is the case for the intrinsic curve. Every exhibited difference of the on-wafer standards derives from an external cause and not from

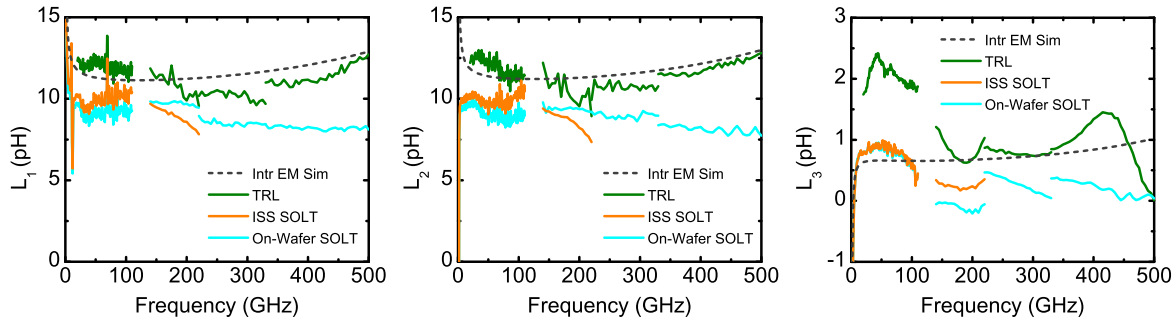


Figure 3.37: All the inductances measurements and intrinsic HFSS simulations of C-S applying different calibration methods (run 2). Note: the reference plane position after TRL calibration (and its intrinsic simulation) and SOLT calibration is the same.

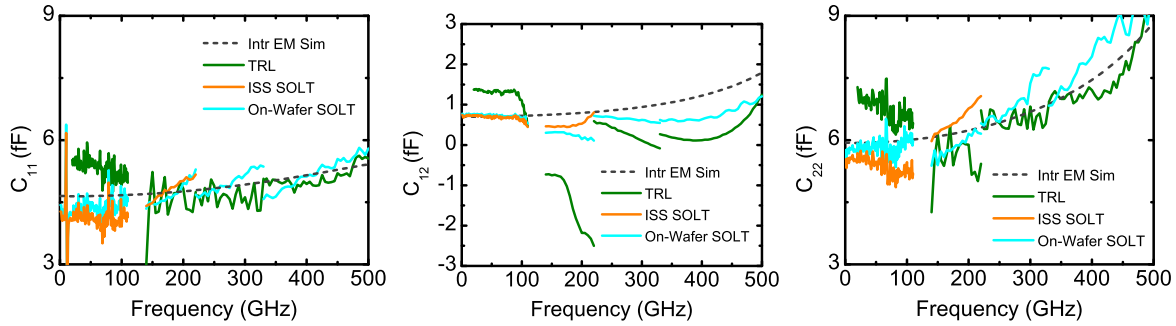


Figure 3.38: All the capacitances measurements and intrinsic HFSS simulations of C-O (no-T0) applying different calibration methods (run 2). Note: the reference plane position after TRL calibration (and its intrinsic simulation) and SOLT calibration is not the same.

the applied algorithm, such as different on-probe contact positioning, weary of test structures' materials, aging of test fixture, etc... The probe setup and the raw data sets are in fact the same.

Finally, the capacitances of C-O (no-T0) are studied (Fig. 3.38). In this case, ISS SOLT provides quite acceptable results up to the second considered band, and on-wafer SOLT proves to follow pretty much the same trend of ISS SOLT at LF, again, and the intrinsic curve in general, up to the last band, although it somewhat lacks of band continuity, compared to TRL, showing a stronger effect of parasitics in each band, in this case. Fig. 3.39 reports all the capacitances in a single plot, not only including the intrinsic simulation but also a complete simulation with RF probes up to 500 GHz. The calibrated trends are perfectly replicated by simulation with very few deviation, localized at LF only.

It is however interesting to consider C_{12}: the nonphysical values in the 140-220 GHz band displayed by the TRL-calibrated curve will be discussed in the following. We recall that the probes, as well as the measurement environment and even the data sets used in this range are the same of the on-wafer SOLT which, in turn, yield a completely flat and physical post-calibrated curve. The artifact of the TRL-calibrated C_{12} in this portion of the spectrum is therefore unequivocally not linked to the design of the structure or any bad user manipulation.

In conclusion, we can state that, overall, the TRL calibration as it is performed here (i.e. by setting the reference plane just after the pads), and the on-wafer SOLT, both perform well. Sometimes, and to a certain extend, on-wafer SOLT even seems to outperform TRL in the whole range from 1 to 500 GHz. The exploitation of ISS SOLT has been stopped at 220 GHz since the environment where the DUT lies is inappropriate and, alongside quality of measurement issues, contributed to excessive miscalculation of the error terms. Due to the fact that de-embeding is performed after on-wafer SOLT, on the other hand, the removal of the parasitics is more complete. The good quality of the final trends of on-wafer SOLT calibrated curves highlights that the lumped nature of the standards is sufficiently well captured in the first band by the constant terms inserted

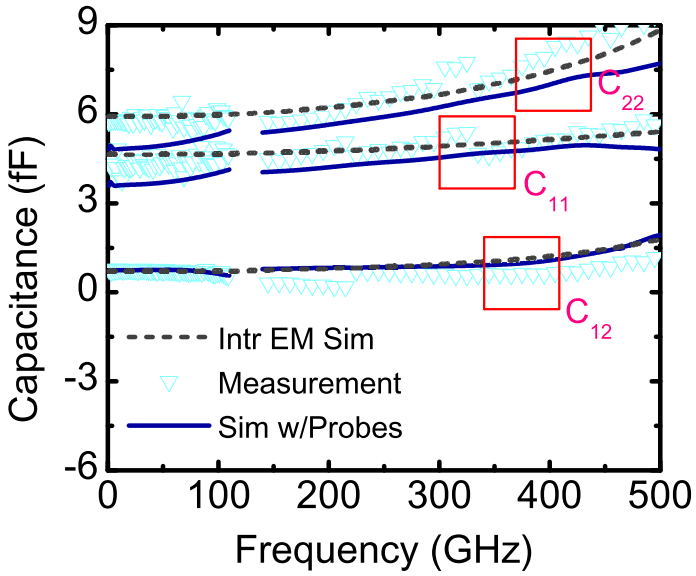


Figure 3.39: All the capacitances measurements, intrinsic HFSS and complete structure simulations of C-O (no-T0) all with the on-wafer SOLT calibration. The reference plane is set at the edge of the P-O access lines.

into the SOLT algorithm –sufficiently enough to deliver proper results. The key point is to test this calibration on transistors, which is done below.

Applying the two calibrations to actual transistor measurements up to 500 GHz yields the graphs displayed in Fig. 3.40. For the transistor analysis, we show here some of the figures of merit that characterize the behavior of the HBTs: the current gain $|H_{21}|$ and the (square root of the) unilateral power gain U, both shown in dB (where we can verify the −20 dB/dec roll-off). They base the calculation of the cutoff frequency f_T and maximum oscillation frequency f_{max}, respectively, here shown including the bias point where the peak value of f_T is reached (i.e. $V_{BE} = 0.9$ V). Raw data come from the same transistor and de-embedding is applied to both calibration. For the active device, the following correction steps are taken from raw data:

- active DUT's raw data ⟶ on-wafer TRL calibration + Z_0 correction ⟹ C-S de-embedding ⟶ C-O de-embedding ⟶ TRL-calibrated active DUT (ref. plane: transistor's contacts at M1);
- active DUT's raw data ⟶ on-wafer SOLT calibration ⟶ P-O de-embedding ⟶ C-S de-embedding ⟶ C-O de-embedding ⟹ SOLT-calibrated active DUT (ref. plane: transistor's contacts at M1).

We remark very similar curve tendencies at LF, with perfect superposition below 67 GHz. Through most of the spectrum on f_T, both calibrations work exactly in the same way as well. On f_{max}, SOLT does not reach the same values (also expected by the HICUM simulation) as TRL. Lower curves on f_{max} in the subsequent bands (at all bias points) might highlight the presence of non-removed parasitic effects. A different (more complex) de-embedding technique may work around the problem, at the cost of complexity. On the last band, SOLT increases at the f_T-peak on f_T, and also outperforms TRL on f_{max}: the results yielded by TRL indicate similar unexpected (and non physical) behaviors as seen on attenuation constant.

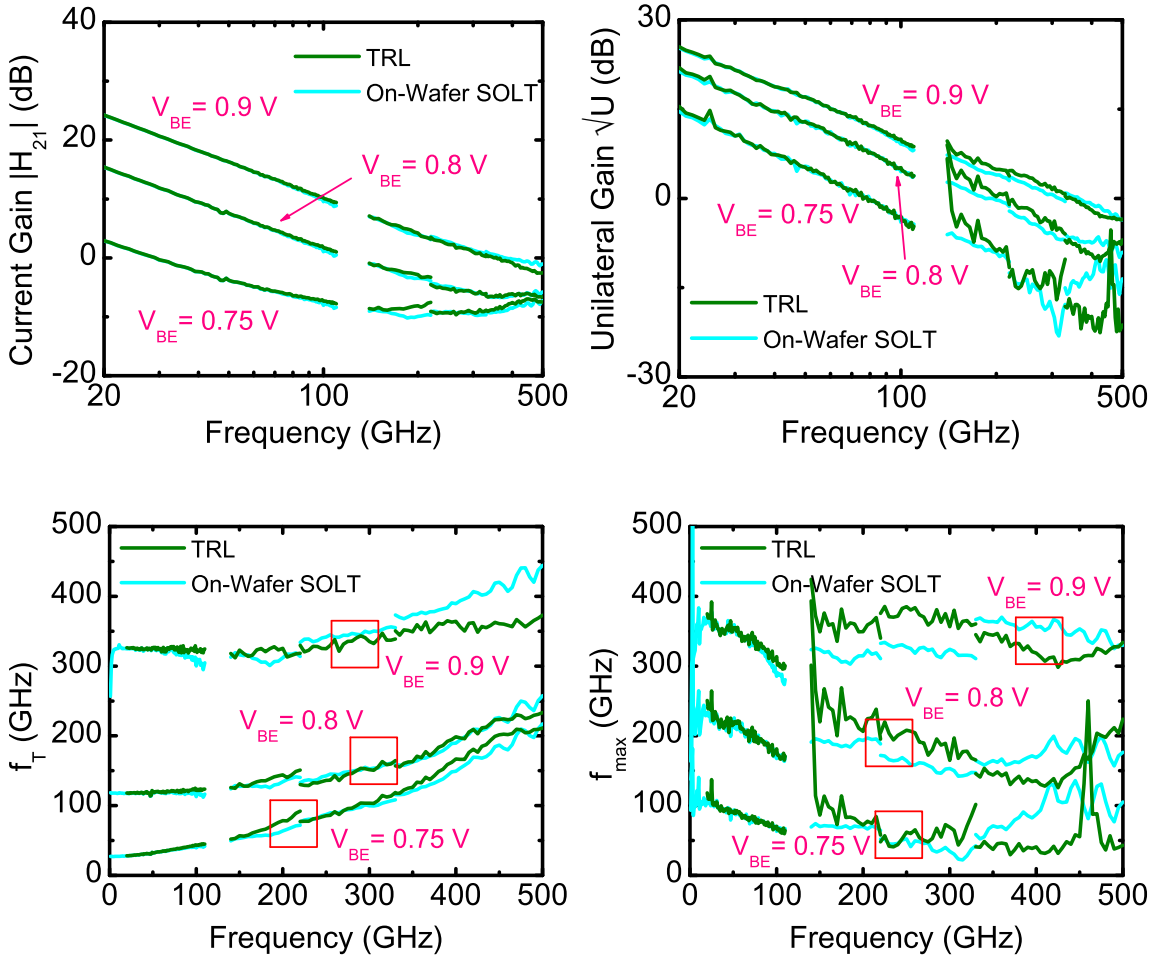


Figure 3.40: Main figures of merit of the HBT measured for different bias points (V_{CB} = 0V, V_{BE} = 0.75, 0.8, 0.9 V).

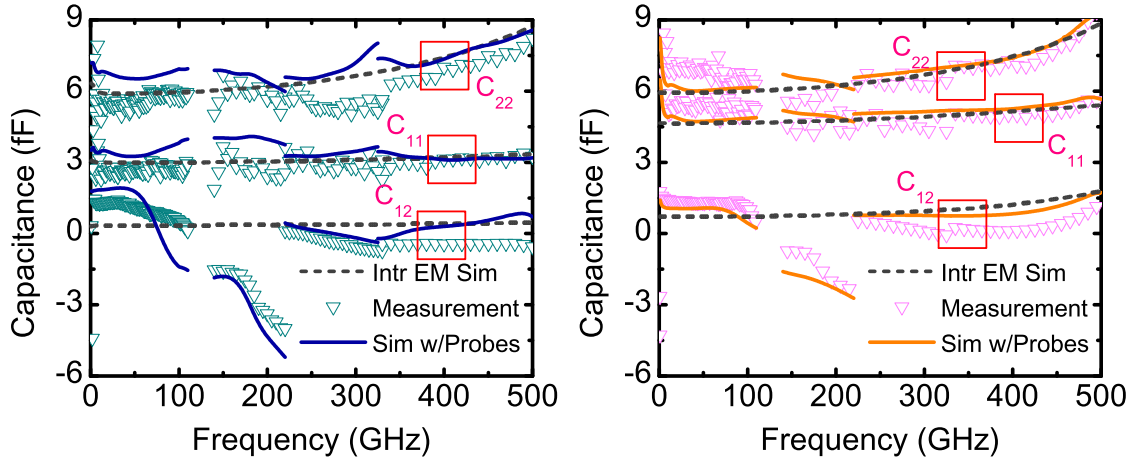


Figure 3.41: All the capacitances measurements, intrinsic HFSS and complete structure simulations of C-O for both run 1 and 2 (no-T0). These capacitances are related to port 1 (C_{11}), port 2 (C_{22}), and the coupling capacitance between the port 1 and port 2 (C_{12}). Data have been measured and simulated with the following set of probes: PP-110, PP-220, PP-330, PP-500.

3.5 Layout Improvement of Run 2

This final section will discuss the effects of the novel run 2 layout design compared to the previous production run. We recall that comparing the verification DUTs and transistor between the runs must be done with caution, since we are dealing with structures conceived with the same purposes, but with differences limited not only to their environment (neighbors placement, pad shapes, geometry difference, etc...) but also to the BEOL shape and size, not to mention hardship in keeping consistency among several measurement campaigns.

3.5.1 Complete-Open and HBT Characterization

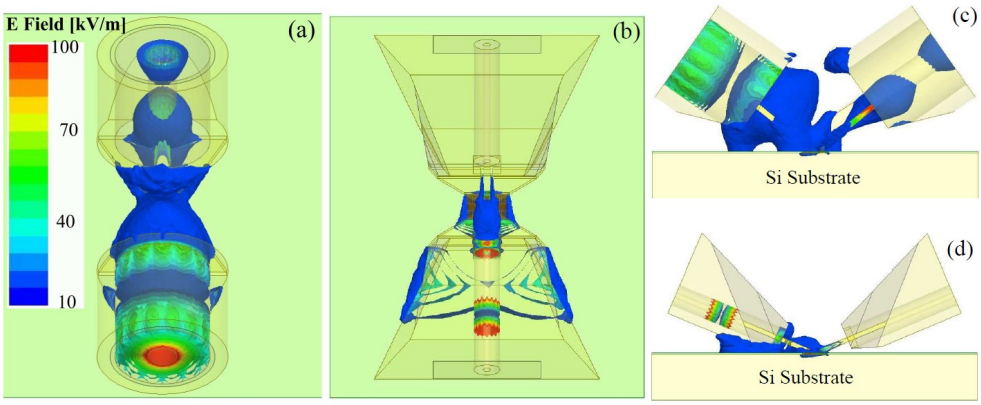


Figure 3.42: E-field distribution (top and side views) at 220 GHz on the C-O using two probe models with very different topologies: PP-220 ((a) and (c)) and PP-330 ((b) and (d)). The field signature is completely different (from [144]).

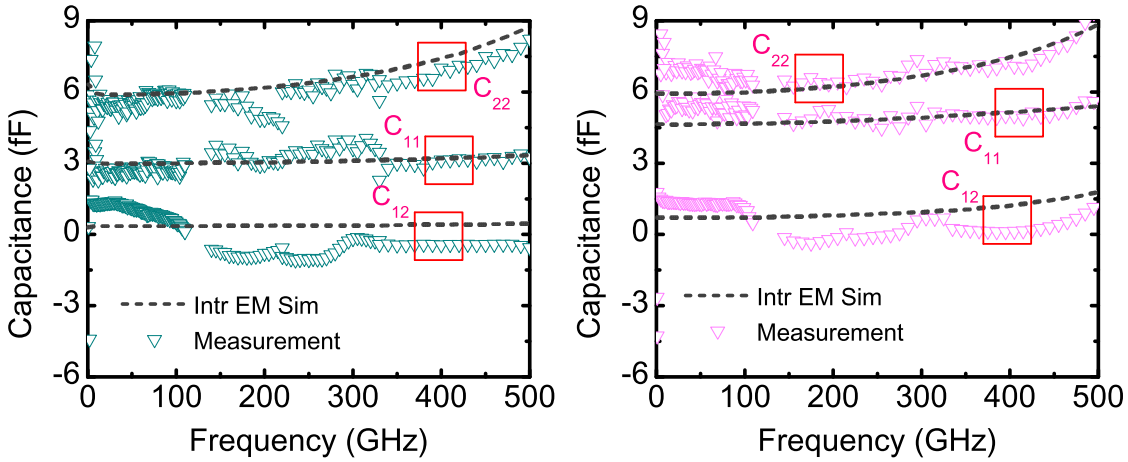


Figure 3.43: All the capacitances measurements and intrinsic HFSS simulations of C-O for both run 1 and 2 (no-T0). Data have been measured with the following set of probes: PP-110, IP-220, IP-330, PP-500.

In Fig. 3.41, all the capacitances of C-O are shown for both run 1 and 2. The measurements are very well replicated by simulation. The asymmetry of C_{11} with respect to C_{22} is due to the different layout designs of the base and collector BEOL stack. It is worth noticing that particularly at very high frequency, the measurements and the corresponding simulations with probes follow the intrinsic simulation, thus confirming the efficiency of the on-wafer TRL calibration. Overall, the run 2 measurements appear to be more consistent and have very good band-to-band continuity, except C_{12}; the lower capacitance at port 1, which is, for that matter, predicted by the intrinsic

simulation, is due to the smaller metallic M1 layer surface and the higher distance between the base metal contact and ground.

In the 140-220 GHz band, though, both the measurement and the simulation with probes diverge from the trend indicated by the intrinsic simulation, particularly the port-to-port capacitance (C_{12}), which becomes clearly nonphysical in both cases (run 1 and 2). As we have just concluded by comparison with SOLT-calibrated data, this is not due to a bad contact or repeatability issues of measurements, and the comparison between the runs confirms once again that the problem does not rise from the design of the test structures, nor that of the RF pads, and not even any neighboring effect. Unfortunately, we cannot tell that the new run design solves this issue, either. This effect may be traced back to the port 1 to port 2 crosstalk between the two RF probes due to their design [144] (see Fig. 3.42). It has been showed that a coplanar probe with a much simpler design does not present this behavior in this band [142].

Elsewhere, we can clearly see that values of C-O turn out to be generally positive for run 2 (except in the 140-220 GHz range), while run 1 measured values become negative from 240 GHz on already. This more physical behavior with respect to run 1 can be definitely attributed to the different configuration of each of the BEOL layers and the continuous ground plane, which provides a different environment around the DUT, more distance from the neighboring structures and less coupling through the pad shield. We turn to a different set of measurements in Fig. 3.43. Here, the probes used in the 140-330 GHz bands are replaced with different sets (IP-220 and IP-330: we are unable to provide the complete EM simulation since these probes' models have not been designed yet). Curves are flatter and oscillate less in run 2, where band-to-band continuity is restored in C_{11}, C_{22} and partially in C_{12} too, which gets fully positive values.

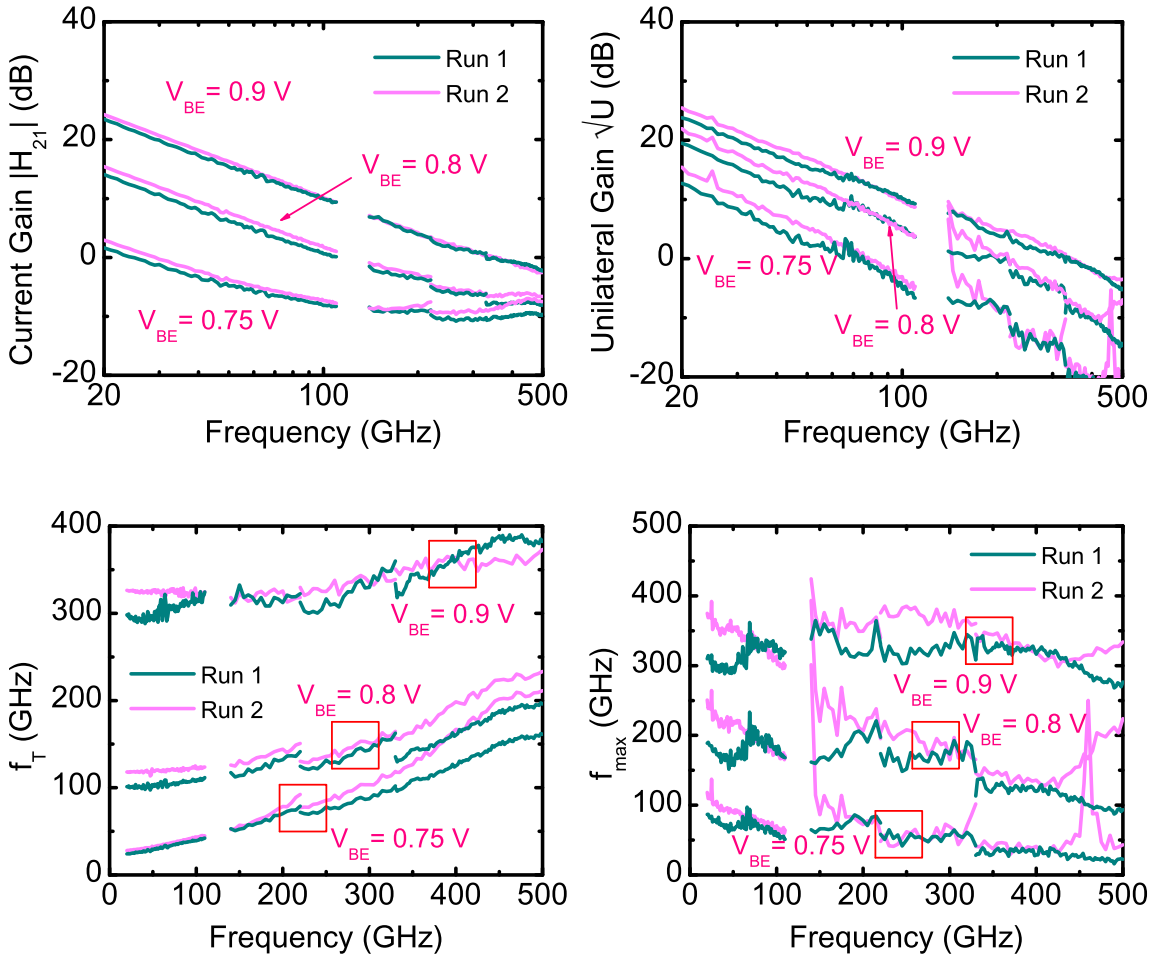


Figure 3.44: Main figures of merit of the HBT measured for different bias points (V_{CB} = 0V, V_{BE} = 0.75, 0.8, 0.9 V).

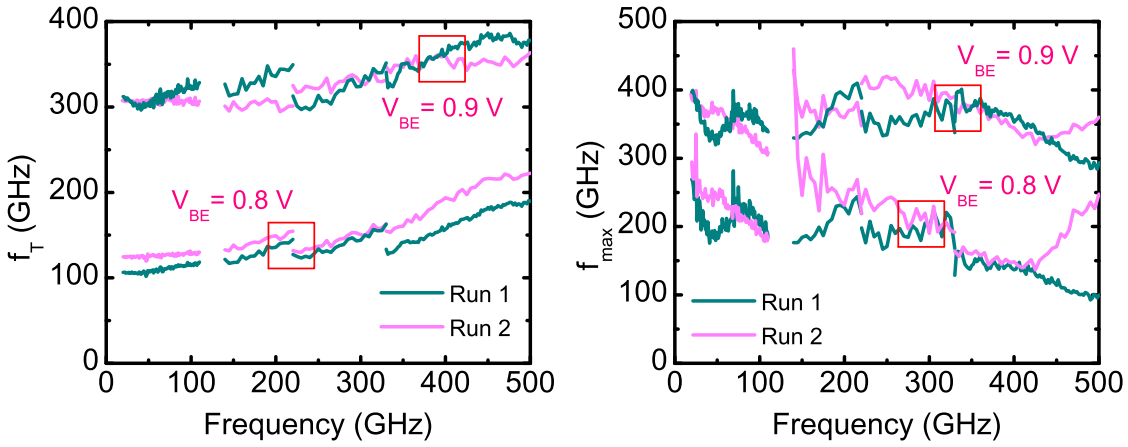


Figure 3.45: Transit frequency and maximum oscillation frequency of the HBT measured for different bias points ($V_{CB} = 0.5\,V$, $V_{BE} = 0.8, 0.9\,V$).

For the sake of completeness, here are the direct comparison of TRL calibration + de-embedding (C-S/C-O) with HBT and test structures from the two different runs, at different bias points. The probes setups are consistently the same for all the displayed measurements: PP-110, PP-200, PP-330, PP-500. In Fig. 3.44 and Fig. 3.45 we present the same transistor biased at $V_{CB} = 0\,V$ and $0.5\,V$, respectively. Please note that the device we are treating, the HBT T-0, comes from the same technology.

Concerning the transit frequency, we can state that, in general, band continuity is better preserved on run 2. An increase of f_T can be observed in the upper frequency bands; in fact, for the lowest frequency range the single-pole approximation from the small-signal equivalent circuit is valid, resulting in a constant gain-bandwidth product. However, the steady frequency increase is weaker for low bias values of V_{BE} on run 1: since we did not point out any bias discontinuity, we believe this is related to an inadequate test port output power at HF, diverging from the power set to –30 dBm at the test ports. The fact that run 2 measurement are more constant and stable, particularly at LF, allows to clearly identify the performance of the HBT.

We examine now the maximum oscillation frequency f_{max}, which is derived from the Mason's gain U and in turn from all the Y parameters. Firstly it is worth noting that Mason's gain is greatly sensitive to parasitic effects and losses, as we have already proved by removing α in the Z_0 correction.

Let us consider the plots in Fig. 3.44. Here, the roll-off of f_{max} in the 1-110 GHz band highlights the presence of losses, even though run 1 exhibits it in a different way (we verified that in both runs, the uncorrected curves share similar trends). An incorrect power delivery to the VNA prior to the used of extenders (below 67 GHz) during the run 1 transistor characterization may justify those distinct losses.

From 140 to 330 GHz, however, f_{max} remains rather constant, for both run 1 and 2. Nevertheless, we observe a stronger noise in the 140-220 GHz band. This difference comes from the fact that we used the measurement from PP-220 in the case of run 2, while IP-220 has been used for run 1. As already pointed out, the probe-to-probe cross-talk in PP-220 has a noise-like perturbing effect on the open capacitances of C-O (but also on the inductances of C-S, see Fig. 3.37), engendering a bad correction.

Run 1 curves perform very closely to run 2 at the beginning of the last frequency range, and while the former do not perform ideally above 420-430 GHz, run 2 curves present, as for them, distinctive trends that are not expected nor legitimate; we observe, for example, a spike at HF for $V_{BE} = 0.75\,V$ and in general, we note that the curves rise starting from 430 GHz. We can try and assume a wear of the contacts and/or inadequate power delivered to the device. The comparison with other calibration standards in the next chapter will help clarify these HF discrepancies. For the trends of run 1, we mention as possible explanation the emitter contact BEOL stack going up and down again to ground. Results on Fig. 3.45 are different, although we can draw similar

conclusions.

3.5.2 Altered Test Structures and Neighbors' Effect

A limitation for the use of TRL is, in real-world applications, the location of structures side-by-side on a die to reduce silicon consumption, with insufficient space to isolate them electro- magnetically. The effect of coupling due to the adjacent structures on the measurements of the DUT (particularly of those located below the RF probes [3, 77]) might be therefore non-negligible, resulting in artifacts such as dips and oscillations on the calibrated S-parameters of the DUT. Thus, it is crucial to reduce the impact of such coupling as well as the probe-to-probe cross-talk [80, 141]. Conclusion from works on the subject [3, 77, 102, 79] relate indeed the loss in measuring accuracy to the probe-to-substrate coupling (where the substrate is, however, made of a semiconductor or dielectric, not of a metallization volume) more than the substrate-to-substrate coupling and recommend a chessboard configuration and increased inter-structure distance, even though for this latter condition a trade-off must be found to contain production costs. They finally recommend homogeneity in the choice of the material (similar dielectric constants to avoid different modes to propagate within the layered dielectric substrate).

In particular, [102] analysed the impact to the EM field on adjacent structures by placing the same DUT with different neighbors and inter-structure distances. They proved that the unconfined field couples more strongly with structures under the shade of the probes and that stray fields increases with frequency and involve structures further and further away from the DUT. The shielding structure designed in our run 2 aims to convey the field in the proper direction and minimize stray energy flux. Our first production run, in fact, did not perform well in terms of port coupling (probe-to-substrate and probe-to-neighbors, e.g. see the "oscillating" trends in Fig. 3.41). Phung *et al.* [77] suggest a longer signal needle and a "sideways shift of successive neighbors" (chessboard) with a wide ground width of a thin-film multilayer microstrip to reduce coupling, and in [79] state that "even for a [CPW] configuration without neighboring line structures, the TRL calibration cannot completely compensate the probe influence", although "the excitation of the resonance in an in-line neighbor [may be] responsible for a dip behavior" which even the absorber material inside the probes is not sufficient to remove. They conclude that the designer should "keep the region of the probe shadow free of structures" [77].

Hence, motivated to investigate the influence of the DUT environment, we consider altered versions of the reference run 2 layout. By this, we mean to alter the run 2 design to match the default run 1 design by introducing spot modifications in order to eventually understand and compare the layout impact, by looking at each contribution individually. We restrict to study the capacitances in the following configurations:

- "Single": In Fig. 3.46a-3.46c, all the structures (DUT and calibration standard thru, P-O and L-110G/L-500G) have first been simulated as isolated structures laying on planes emulating infinite dielectric layers and ground. The DUT, C-O, is subsequently calibrated with the corresponding isolated calibration standards. We consider two different configurations:
  - "Ref": the reference, where each run 2 isolated structures is modelled;
  - "Ox ring": the altered version, where a 10-µm SiO_2 ring surrounds each isolated structure, much like in run 1, and the pad shield is removed.
- "Neighbors" (for short, "Neigh"): In Fig. 3.46d-3.46f, all the structures have been simulated with their actual corresponding adjacent structures. Only the closest neighbors are taken into account, as the electromagnetic impact of more distant structures on the DUT is negligible. For example, the thru used for the TRL calibration (located at position C3 in Table 3.2) has been simulated with all the structures at Bi, Ci, Di, with $i = 1,2,...5$). Where neighbors are not present, copper is placed. The pad-to-pad distance between each DUT and its neighbors is brought to the same as run 1 (45 µm). Three different configurations are considered for these plots too:

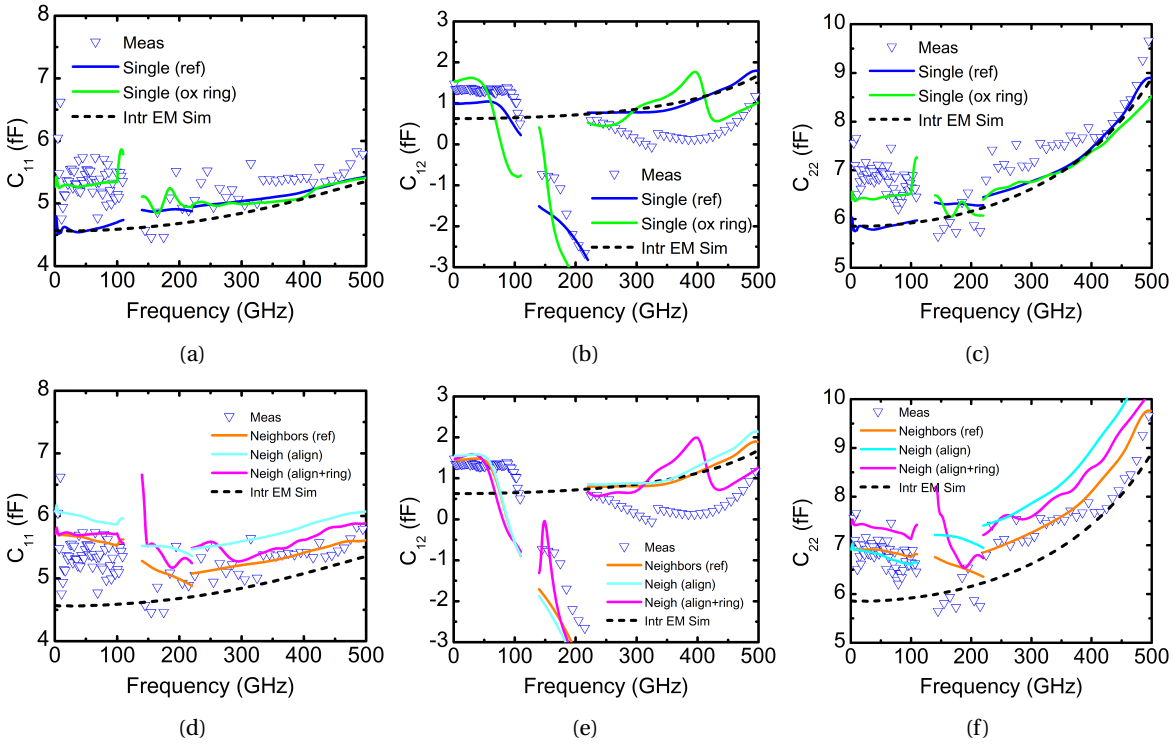


Figure 3.46: Run 2 measurement vs. simulation. Single structure with and without SiO_2 ring (from (a) to (c)), DUT with neighbors (adjacent structures, from (d) to (f)) in the actual chess-board configuration, with aligned neighbors, and in a "pseudo run 1" configuration (aligned structures with a 10 µm SiO_2 ring around them).

- "Ref": the reference, where every run 2 structure is not surrounded by any ideal infinite plane of dielectric and metal, but by its actual neighbors, in their actual position, as they are on the wafer. Indeed, Phung *et al.* put the stress on the measurement degradation produced by different probe positions and location on the wafer, varying structures in the neighborhood even when structures are completely symmetric [78];
- "Align": the first altered version, where the neighbors are also present but do not present the same chessboard configuration: they are aligned in columns and rows, instead. For simplicity, the common chosen neighbor is only one: pad-open, since this structure has proved to be more prone to EM coupling, thus representing a "worst case" situation;
- "Align+ring": in this second altered version, neighbors are also aligned and surrounded by P-O, and each test structure has an oxide ring around it, like in run 1: therefore, this represents a "pseudo-run 1" case.

From Fig. 3.46a-3.46c, the effect of the oxide ring is visible. Overall, the curves show small fluctuations around the reference (mostly on C_{12}), and larger deviation from the intrinsic curves at low frequency. In Fig. 3.46d-3.46f, the contribution of neighbors is also considered. As we can see, the curves representing the reference case with neighbors, the orange curves in Fig. 3.46d-3.46f, keep close to measurements, as expected, but the curves in the case with no neighbors (blue curves in Fig. 3.46a-3.46c), do not deviate considerably: in fact, they differ less than 1 fF all over the spectrum, except in the 1 – 110 GHz band, where deviation slightly exceeds this value. This important result yields to the conclusion that by the optimized design of run 2, the impact of the neighbors is considerably reduced. Also, this fact makes us confident on the use of "single" structures instead of models with adjacent structures around them for EM simulations, which greatly reduces simulation time and complexity.

The alteration of the neighboring environment specified in Fig. 3.46d-3.46f by the cyan and

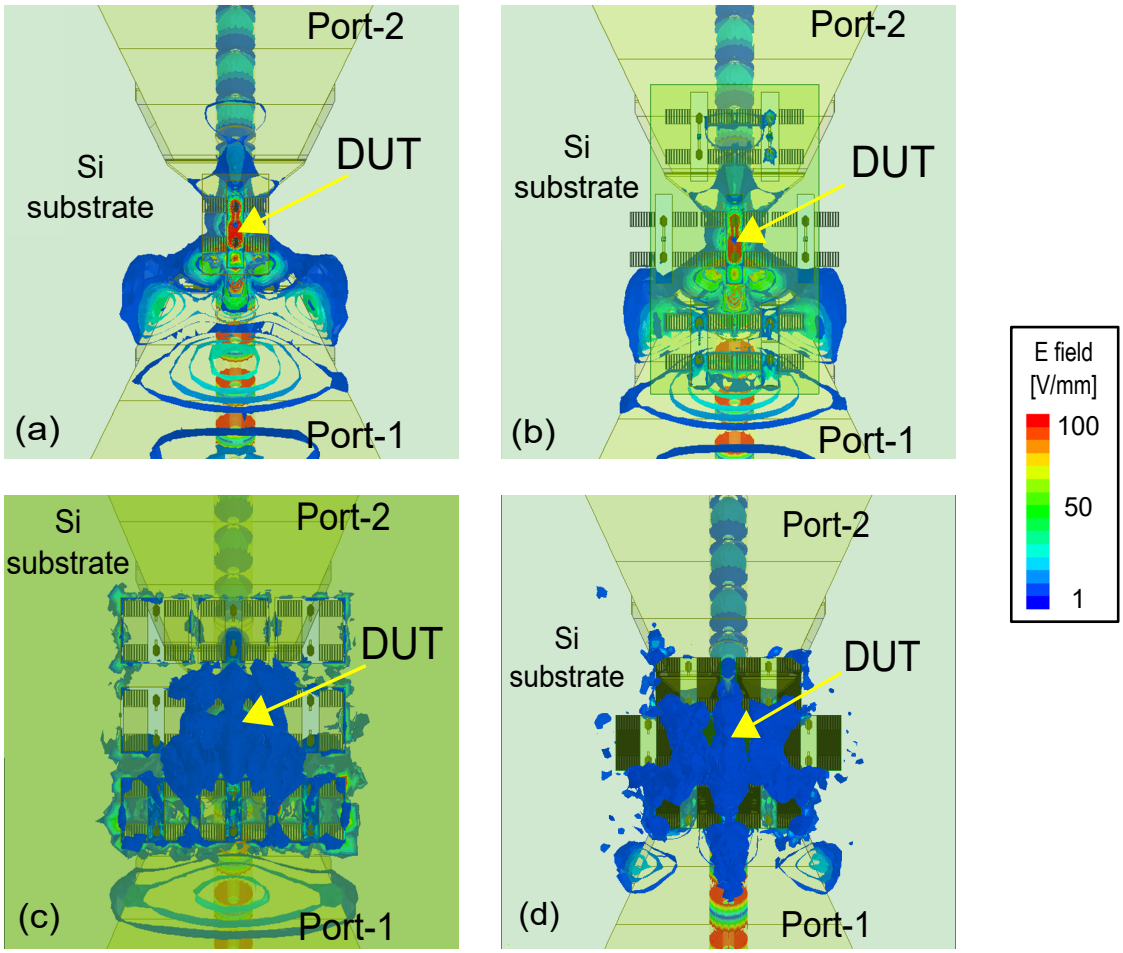


Figure 3.47: HFSS-simulated electric field contour (back view) for DUT complete-open at 500 GHz: single structure (a), neighbors (b), aligned neighbors with oxide ring (c), and neighbors in a more dense configuration (no continuous ground plane) (d).

purple curves also provides some interesting insights. If we first consider the case where the structures have just been aligned ("align"), capacitances C_{11} and C_{22} keep above the reference configuration ("ref") in all the considered frequency spectrum; the coupling is reduced thanks to the pad shielding and the chessboard configuration. Note that, on the other hand, C_{12} is only very moderately affected. The addition of the oxide ring ("align+ring") disturbs the trends of the port capacitances generating a small ripple, but more strongly the trend of C_{12}, which is similar in shape to C_{12} in the case of a single structure with oxide ring (Fig. 3.46a-3.46c). Therefore, on C_{12} the effect of the alignment is negligible compared to the one produced by the oxide ring while, on the contrary, closer and aligned neighbors affect more the capacitances between the ports and the substrate, with an additional capacitance adding up to the one which is already present (offset on C_{11} and C_{22}).

We observe now the electric field 3D contour obtained by HFSS. In the backside view of the DUT and its neighbors (Fig. 3.47), we see the E-field concentrating on the DUT and beneath the probe corresponding to port 1, where the field is excited, and the most intense E-field is efficiently confined inside the space created by the pad shield. The field contour has the same shape either with or without the presence of adjacent structures (Fig. 3.47a, Fig. 3.47b). When the dielectric ring is present (Fig. 3.47c), the E-field is heavily affected. The intensity of the field increases around the DUT and below the excitation probe and we can clearly see the field densifying around every adjacent structure. In Fig. 3.48, the penetration inside the silicon substrate is well depicted. Where no oxide ring is present (Fig. 3.48a and Fig. 3.48b), the E-field is only slightly visible in the substrate region below the DUT. Also, no difference can be noticed in the field shape when the neighbors are

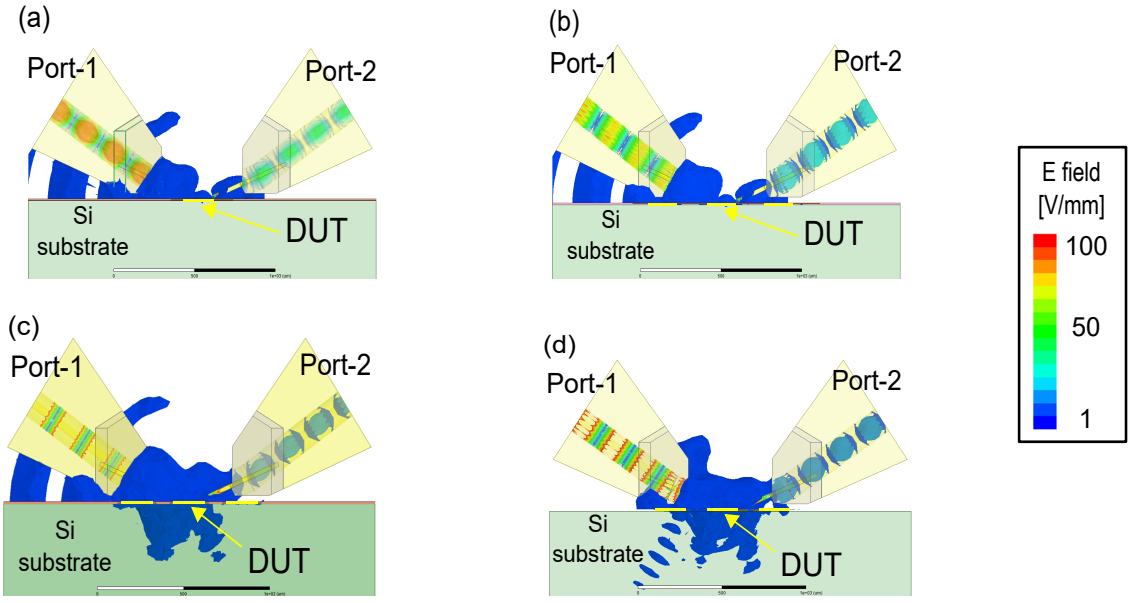


Figure 3.48: HFSS-simulated electric field contour (side view) for DUT complete-open at 500 GHz: single structure (a), neighbors (b), aligned neighbors with oxide ring (c), and neighbors in a more dense configuration (no continuous ground plane) (d).

considered, thanks to the continuous ground plane. On the other hand, when aligned neighbors are in place (Fig. 3.48c), the field is free to couple below the DUT with all the neighbors through the rings.

Fig. 3.47d and Fig. 3.48d introduce yet another configuration. It aims to represent a more dense layout design approach of test structures: the neighbors in this case are in chessboard configuration but no continuous ground plane is present. In fact, no metal volume is connecting the structures, which are only surrounded by the dielectric. Also, their mutual distance is even more reduced, in order to create the most compact design: the area occupancy is reduced by a quarter. The electric field in Fig. 3.47d consequently scatters and permeates all the neighboring structures more intensely than in Fig. 3.47c. Fig. 3.48d, on the other hand, shows that the probe-to-substrate coupling is reduced (since no metal is under the probe), but the field uncontrollably propagates through the lossy silicon substrate; however, while the coupling can be removed by calibration, the dispersion of the field in the substrate is harder to correct, making it more prone to measurement errors.

In conclusion, we see that the shielding structure, the chessboard placement and the removal of the dielectric ring result in a more confined energy flow and avoid all artefacts on any capacitance, as confirmed by simulation of the altered run 2 layout. Also, the electric field contour reinforces our motivation in the use of a continuous ground plane, whereas a widespread approach employing a dielectric substrate leads to uncontrolled coupling with neighbors and the substrate. Our layout completely avoids any structure-to-structure coupling, but particularly makes the probe-to-substrate coupling, that, as mentioned by, e.g., [3, 77], is the main cause of the corruption of measurements, more controllable and easy to remove, thanks to the metallization provided by the continuous ground plane.

Chapter 4

Evaluation of Innovative Calibration Standards' Design

Contents

IN THE PREVIOUS chapter we evaluated the performance of SOLT and TRL calibrations on different test structures for microwave measurements and presented the new layout of our structures, evaluating the improvements they bring in terms of electromagnetic isolation and optimization for high frequency characterization and parametric extraction.

We employ in the following the production run 2 that we have presented and we introduce the remaining test structures present on the die. Thanks to them, alternative approaches to TRL calibration are suggested below, to allow new forms of on-wafer TRL calibration.

The first set of test structures attempts to perform one-tier calibration that places the reference plane directly in proximity to the transistor without resorting to the use of de-embedding, a source of additional errors which complicates the parametric extraction. This technique is based on the drawing of microstrip lines and other calibration standards at the level of metal 3.

A second design that is studied is based on meander lines, and allows to perform a calibration using lines of different length, even if keeping the inter-probe distance constant, and therefore avoiding to distance the probes from each other during the measurement campaign.

We are also taking a glimpse at other design attempts that have not yet been fully characterized, most notably one-tier calibration by microstrip lines drawn directly on metal 1.

4.1 Toward a One-Tier Calibration: the M3 Layout

As we asserted when the concept of de-embedding was introduced early on, the complex geometries connecting the transistor accesses and the RF pads are usually excluded from any calibration routine. Indeed, in a two-tier calibration process, the first step aims to set the reference plane right after the probes tips. The removal of the access and BEOL metal layers is performed by the second-tier calibration where usually a simple two-step approach (open-short) is preferred, and we also adopted in this work.

Of course, however, a plethora of different approaches have been explored in literature to seize the distributed nature of parasitics in the most accurate way, even at high frequency. Most of these methods are based on elaborate lumped circuits.

In [55], Koolen *et al.* presented a "distributed model", assuming that the shunt parasitics are partly distributed over the interconnect lines, an initial attempt of dealing with parasitics generated at HF; another early distributed model for the series parasitics, accounted for by a third dummy –a thru, was made in [19]; later was proposed a complete four-step technique [54] suited for large fixture gaps; an even more recent three-step chain matrix on-wafer de-embedding, which again takes into account probe pad impedances and admittances as well as interconnections [20]; a five-step de-embedding approach to account for all the parasitics from pads to transistor level, through both HF and DC measurements [83]; another five-step de-embedding applied to a CMOS technology [53]; a four-step de-embedding taking into account parallel and series elements of pads, access lines, and top-down via-holes [124]; a 12-term three-standard open–short–load de-embedding for accurate cross-talk evaluation [118], etc...

As a drawback, when the complexity of all these models is augmented, the probability of measurement errors, uncertainties and performing time also grow, as well as the need for well-known extra standards and occupancy of the die surface [125, 25].

An alternative to lumped element-based de-embedding technique is given by the transmission lines and their inner distributed nature, which is theoretically expected to work fine at HF and reduce wafer consumption [147, 61]. Finally, 4-port models have also been proposed alternatively to lumped models, employing multiple dummies and making no assumption at all on the nature of on-wafer parasitics [59, 138].

Rather than a traditional calibration + de-embedding approach, one can imagine to pursue the goal of a one-tier calibration technique using calibration standards only, pushing the reference plane to the DUT terminals without extra de-embedding steps. Any assumption on the lumped nature of the on-wafer structure's fixture is avoided (as for us, it has been modelled by the complete-open and complete-short).

Some showed that the joint contribution of test pads and accesses leading to the HBT could be still approximated as lumped elements up to 170 GHz [145], opening the way to a calibration approach comprising of all the metal connection to the HBT level. Rumiantsev *et al.* explained in [93] that moving the reference plane down to transistor level could remove the biggest portion of parasitics in one step, and demonstrated that a one-tier on-wafer mTRL calibration to the transistor terminals (M1) performed as good as a classic two-tier off-wafer calibration + de-embedding up to 110 GHz. However, their analysis has never been extended up to 500 GHz.

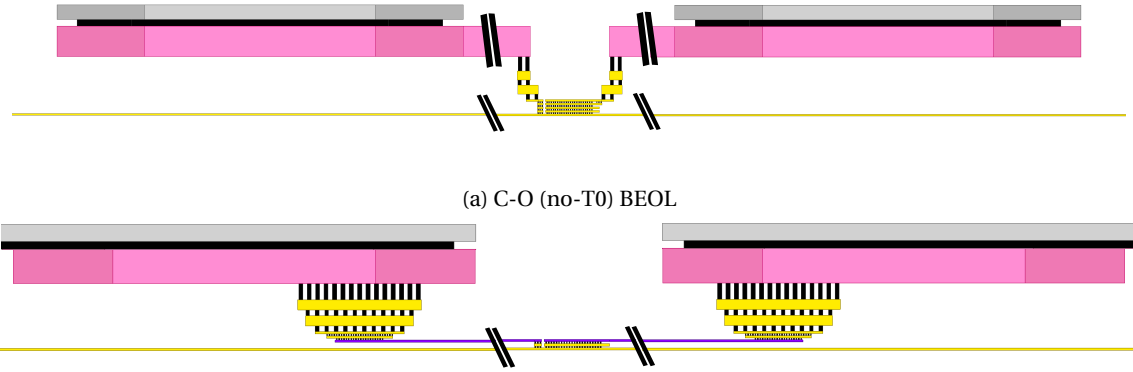
In addition to that, for an accurate and rigorous one-tier TRL, it is necessary to build the access lines as well at the bottom metal level, i.e. at M1. As pointed out by Galatro *et al.* [40], this will necessarily expose the lines to the lossy and poorly controlled substrate, degrading the propagation characteristics of the line. The authors came up with a solution for their CPW transmission lines named capacitively loaded inverted CPW (CL-ICPW). Their idea comes from the observation that the higher permittivity of the substrate (silicon) stores proportionally more energy than that of the oxide, where the BEOL is built. Consequently, the capacitance per unit length of the oxide needs to be increased to compensate this effect. In order to avoid excessive ohmic losses due to a thinner signal line as a consequence of bringing the grounds of the CPW closer to the signal trace, they opted for artificially boosting the dielectric constant of the oxide with perpendicular floating metal bars in close proximity to the transmission line.

In a microstrip-based topology like ours (please note: we focus from now on to our run 2 implementation) we do not suffer from particular dispersive effects related to the silicon substrate, since the ground plane built on M1 divides from the underlying substrate and the holes due to density rules are few and electrically small.

We have decided to draw our transmission lines at an intermediate level, i.e. metal 3, as close as possible to the transistor's accesses. However, this means that we will never be able to de-embed the whole metal stack contribution, since a small connection (V2-M2-V1) is inevitably remaining.

In this first part of the chapter, we are going to study how the impact of such a design, henceforth called "M3 layout" and allowing a "M3-TRL" calibration, affects the final measurement results up to 500 GHz.

4.1.1 M3 TRL Calibration Standards



(a) C-O (no-T0) BEOL

(b) C-O (no-TM3) BEOL

Figure 4.1: Side view artwork of C-O connection to pad for TRL at M3 and M8. The signal pad is shown in grey, M8 is shown in pink, M3 in purple, the other metal layers in yellow and vias in black.

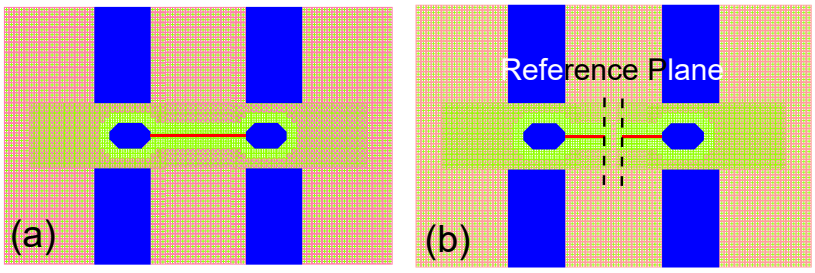


Figure 4.2: Layout view of part of the on-wafer TRL calibration kit. (a) thru (M3), (b) P-O (M3). In pad-open (M3), post-calibration reference planes are shown with black dashed line.

We draw the attention of the reader back to Fig. 3.9 and Table 3.2: in column D, the calibration and de-embedding standards for the M3-TRL are located, whereas the corresponding HBT (T-M3) is at F10.

The standards' properties are very much like the already-presented run 2 structures (we will refer to them as "classic TRL"), and the HBT is exactly the same, except most of the BEOL descends vertically from below the RF pads and connects the access lines which are now at M3. Their dimensions have been consequently modified to provide approximately 50 Ω.

Fig. 4.1 depicts the C-O's BEOL, comparing C-O (no-TM3) and the classic C-O (no-T0). The top view of thru (M3) and P-O (M3) are shown in Fig. 4.2. We can notice that the reference plane after calibration is approached in the vertical direction to M3, thus removing the whole M8-M3 parasitic addition (Fig. 4.1a), as well as in the horizontal direction, since the M3 access lines extend much

	Property	Classic TRL	Modified TRL	M3-TRL
		Value (µm)		
Transmission line geometry	S-G horizontal distance	28.6		
	S-G vertical distance	5.6		0.5
	Ground plane thickness	0.17		
	Line thickness	3		0.2
	Line width	7.7		2
Line lengths (Ref. plane's post-cal position)	Thru length	65	11	15
	L-110G length	595	541	545
	L-500G length	185	131	135

Table 4.1: List of properties for the M3 and M8 lines, considering the different positions of the reference plane. "S-G" indicates the distance between strip and ground. "Modified TRL" is a classic TRL where O-M8 is used as a reflect instead of P-O, and the reference plane position is changed.

further from the signal pads (Fig. 4.2b). Eventually, with this M3 layout, the reference plane is set from 65 µm to 15 µm closer in the horizontal direction and from 5.62 µm to 0.53 µm closer in the vertical. We do this to conventionally locate the reference plane at the edge of the C-O: the one related to the M3 layout. Table 4.1 sums up some of the topological values of the M3 microstrip compared to the M8 microstrip.

Once more, the line dimensions have been designed to carry a quasi-TEM mode: by intrinsic simulation we verified, similarly to what has been done for the classic case, that only the TEM mode propagates with non-negligible intensity in the strip. However, looking at the transition from pads to line in Fig. 4.1a, we can see that the signal, before propagating on the microstrip, is inevitably going to be non-TEM, as it flows through heterogeneous metallic connections (flat metal layers and thin vias). Yet this complex propagation will be included into the calibration error terms, just summing up to the multitude of imperfect probe to planar structure's transitions.

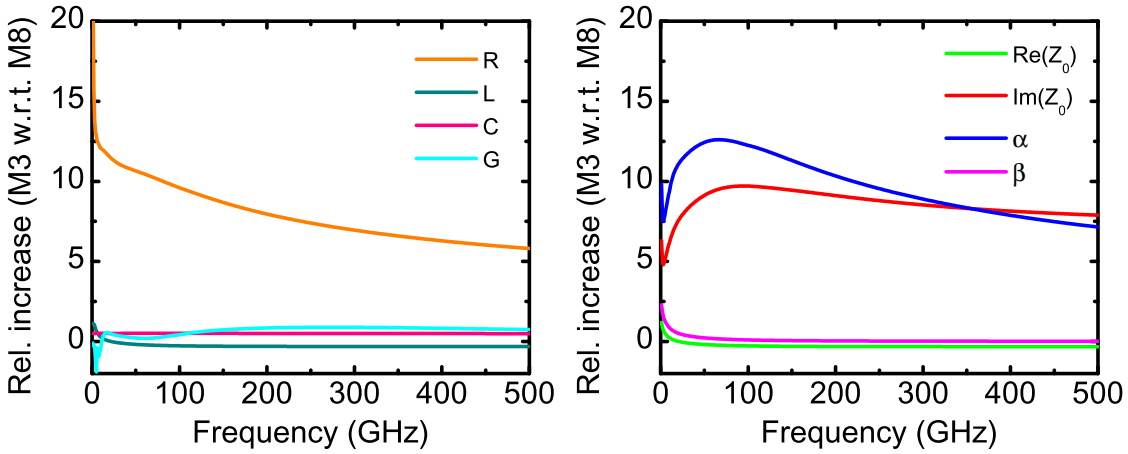


Figure 4.3: Simulated relative increase of the line parameters of L-500G built on M3 compared to the same built on M8.

The simulated line parameters' relative increase are shown in Fig. 4.3. They are extracted from the intrinsic models of the lines: R, L, C and G are calculated using the chain matrix (Eisenstadt) method [26]. We note a strong increase of the line resistance (7 to 13 times higher), while the other parameters remain comparable for the two structures (C and G increase slightly, while L remains overall constant).

Since the material properties vary very few, the physical explanation comes from the shrink in the cross-sectional dimension of the central conductor at M3, that consequently highly increases the associated resistance.

Moreover, in the M8 layout case the losses, both ohmic and conductive, are low and almost

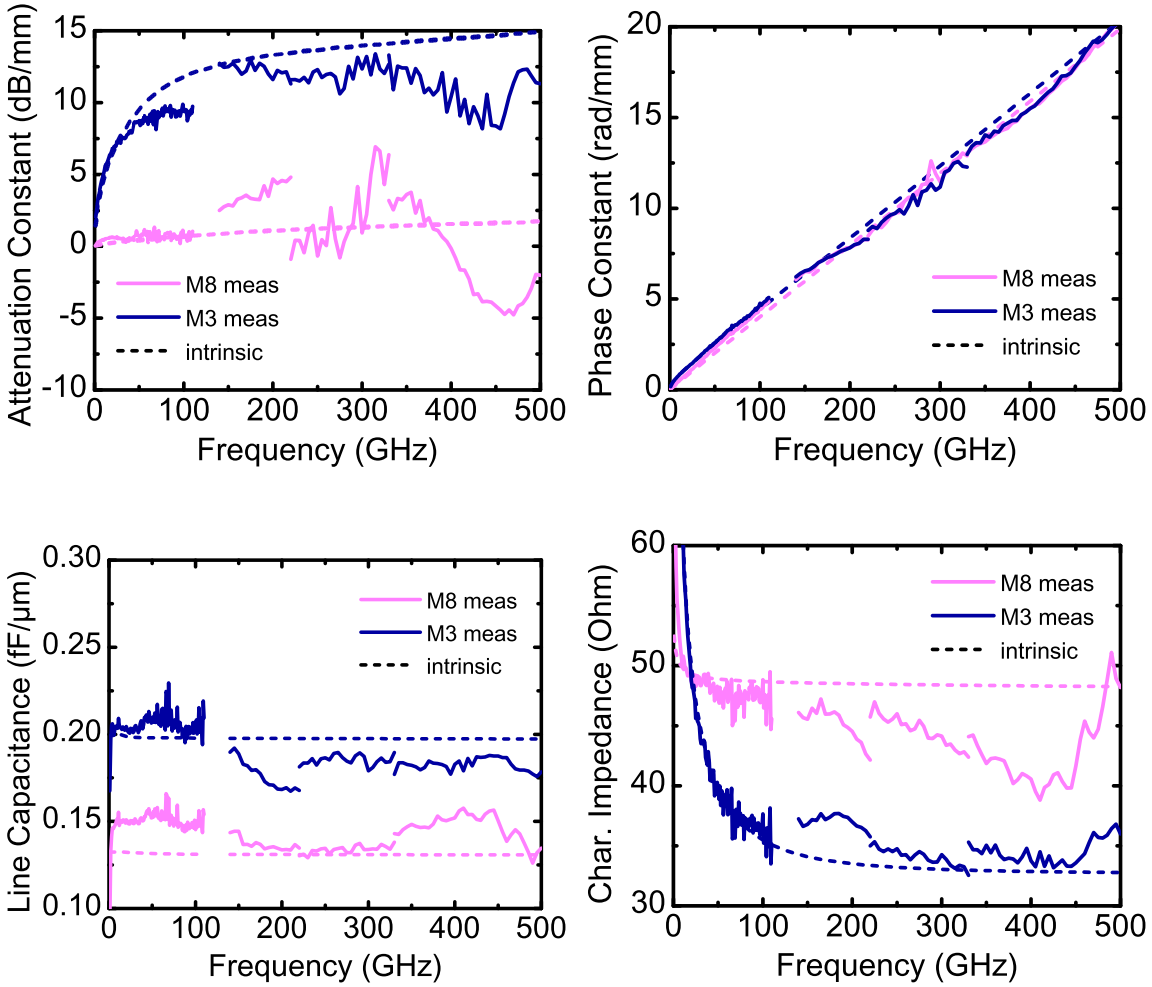


Figure 4.4: Measured electric parameters of L-500G and L-500G (M3), marked as "M8" and "M3", respectively. Comparison with the intrinsic simulation of the same lines.

frequency-independent. The hypothesis of low losses ($R \ll \omega L, G \ll \omega C$) is less likely to hold true on the M3 lines. Here is the complete complex definition of the propagation constant [84]:

$$\gamma = \alpha + j\beta = j\omega\sqrt{LC}\sqrt{1 - j\left(\frac{R}{\omega L} + \frac{G}{\omega C}\right) - \frac{RG}{\omega^2 LC}} \tag{4.1}$$

Indeed, while globally the losses in the M3 line can still be considered low ($RG \ll \omega^2 LC$) and we can neglect the additional $\frac{RG}{\omega^2 LC}$ term, the Taylor series expansion, which, in the case of the M8 line allowed us to simplify the previous equation by:

$$\gamma = \alpha + j\beta \simeq \frac{1}{2}\left(R\sqrt{\frac{C}{L}} + G\sqrt{\frac{L}{C}}\right) + j\omega\sqrt{LC} \tag{4.2}$$

is valid at the first order only above approximately 200 GHz, for the M3 layout, essentially because the condition $R \ll \omega L$ is violated.

In Fig. 4.4, we can indeed see a strong frequency dependence of alpha, and Fig. 4.3 shows that the losses are on average 10 times stronger. We are led to trace them back essentially to higher R, i.e. they are mainly conductor losses. The imaginary part of Z_0 also increases because of the grown R. The phase constant still maintains linear over frequency (Fig. 4.4).

Although relatively small compared to the change of R, we can also observe some non-negligible changes in the measured and simulated line capacitance and characteristic impedance of the line, too. Indeed, for the M3 line, Z_0 is 10-20 Ω lower. Nevertheless, this value is still acceptable and is

corrected by the lumped-load method, which still holds valid, since the capacitance remains overall rather constant and the $G \ll \omega C$ condition does not variate considerably.

It is also worth pointing out that, due to the horizontal displacement of the reference plane, now closer to the center of the thru, a large part of the M3 line attenuation is removed after TRL calibration. Nevertheless, since the attenuation per unit length is still high, the cumulative losses on the longer line (L-110G (M3)) still have considerable repercussions on the calibrated parameters below 110 GHz.

4.1.2 Reflect: Open-M8 and Reference Plane Location

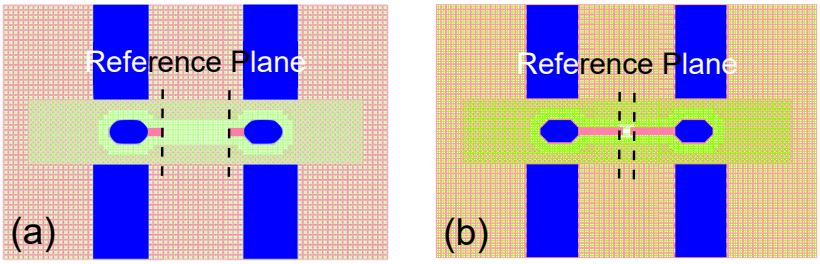


Figure 4.5: Artwork showing the reference plane location for the classic and modified TRL calibration. Located at 15 µm from the signal pads in the classic TRL with reflect = P-0 (a), located at 42 µm from the signal pads in the modified TRL with reflect = O-M8 (b).

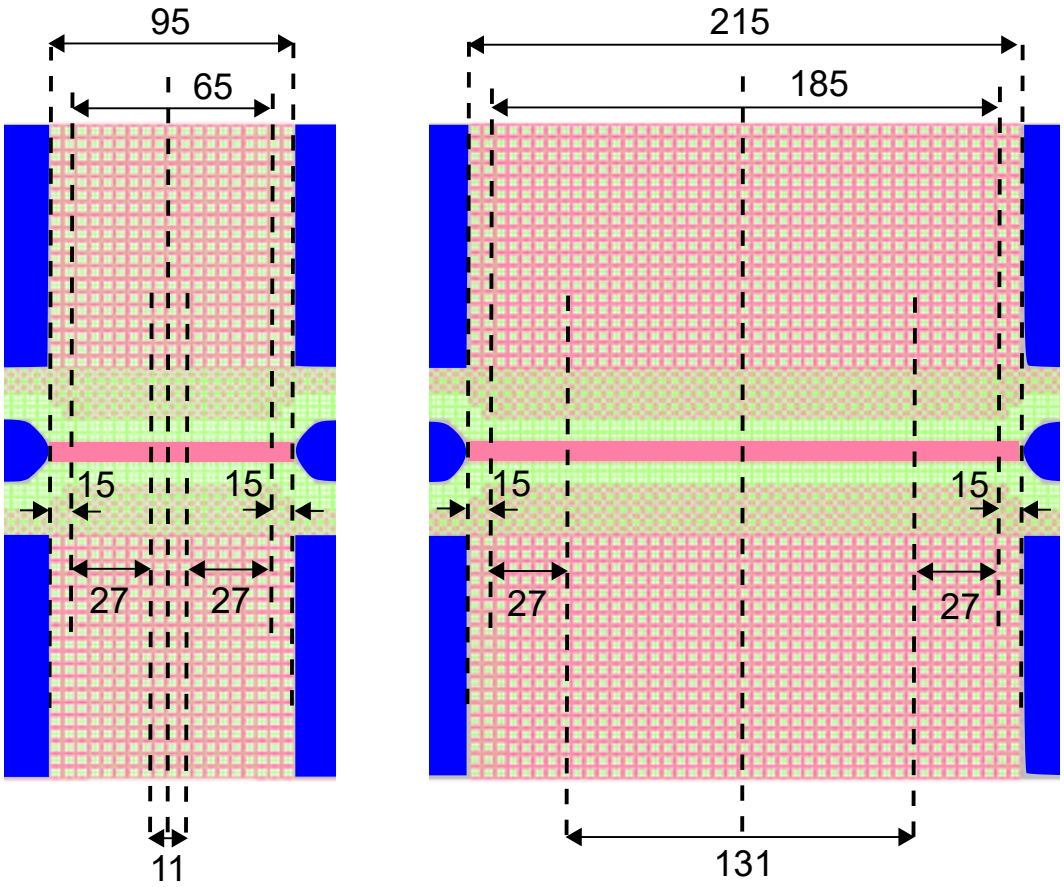


Figure 4.6: Position of the reference plane and associated distances (in µm) for thru (left) and L-500G (right) after classic and modified TRL (reflect = O-M8).

Before starting the evaluation of the M3 layout performance, let us introduce yet another calibration structure. In Fig. 3.9 and Table 3.2, at positions E9 and E11 we can see the open-M8 and

short-M8 structures. These two calibration standards have been designed to provide an alternative TRL reflect to the pad-open or pad-short, and we name such a calibration a "modified TRL".

In Table 4.1, we can see the equivalent lengths of the M8 line after calibration with O-M8 as a reflect: the reference plane, at the edge of the access lines that now extends further away from the pads, is moved 27 µm closer to the DUT at both ports. It is now almost at the same horizontal position as the M3 lines case, even though its location is at top metal level (M8).

See Fig. 4.5 for a comparison between the reference plane for the classic TRL, at the edge of the P-O access lines, and for the modified TRL, at the edge of the O-M8 access lines; see also Fig. 4.6.

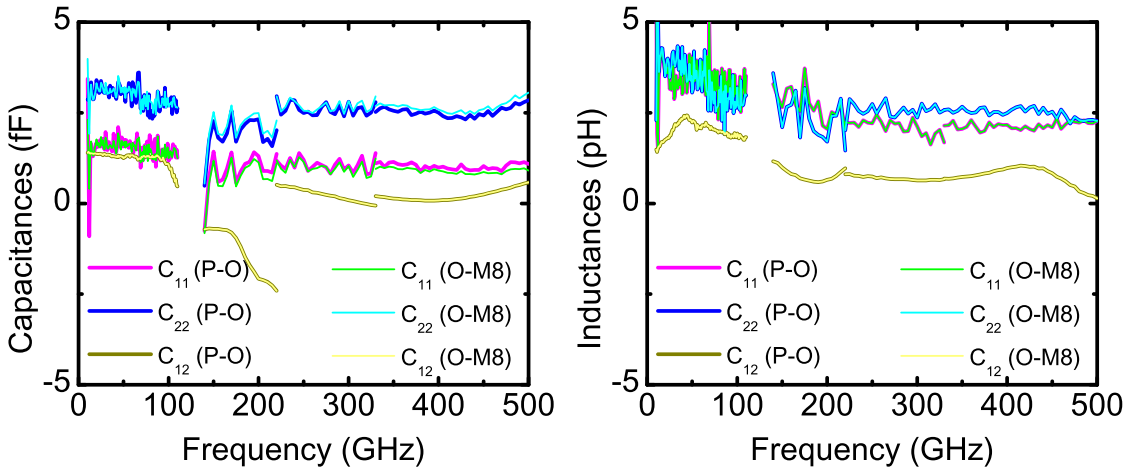


Figure 4.7: C-O (no-T0) capacitances and C-S inductances after calibration. Calibration is performed by classic TRL approach with two different reflect: P-O and O-M8. The reference plane, however, is defined by the algorithm to set the length of thru to 11 µm in both cases.

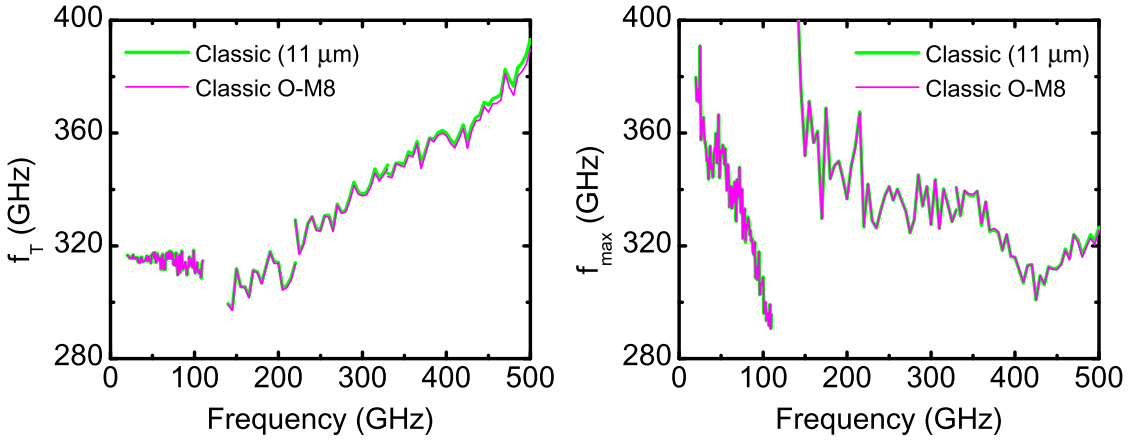


Figure 4.8: Transistor's FoM after two-tier calibration. Calibration is performed by classic TRL approach with two different reflect: P-O (classic) and O-M8 (classic O-M8). The reference plane, however, is defined by the algorithm to set the length of thru to 11 µm in both cases.

Our rationale for the on-wafer HF frequency characterization has always been to set the reference plane where the edge of the reflect standard is located. Firstly, this allows us to avoid any changes in the complex value of its reflection coefficient, due to the influence of longer access lines (e.g. attenuation). This might lead to artefacts on the calibrated results. It is true that reflect standards are not required to be ideal, but we take a conservative approach anyway.

Secondly, reference planes at pad-open's edges also prove to be a more flexible choice when comparing TRL-calibrated DUTs with SOLT-calibrated, since the TRL's reference planes are very close to SOLT's (the RF pads), as well as when evaluating non-conventional calibration standards' designs, such as meander lines (presented in the next section).

We study below results that are put into context by using open-M8 as alternative reflect standard. In fact, yet another practice that we have employed throughout this thesis and our research work in general is to set the reference plane where the edge of the open reflect is. Even though that is a perfectly legitimate choice (the lines are homogeneous before and after the position of the planes), as we have seen in the previous chapter, it is probably not the best choice, since several parasitics remain embedded. The next open-M8 analysis, with the consequent displacement of the reference plane will show to which extend the reference plane closer to the DUT can be beneficial.

Finally, Fig. 4.7 and 4.8 tell us that the choice of the reflect type (either P-O or O-M8) is completely irrelevant in terms of both one-tier calibration and even after applying the extra de-embedding correction, as curves are almost perfectly superimposed at every frequency. Consequently, the location of the reference plane coinciding with the edges of the reflect open is an entirely conventional rule.

4.1.3 Calibration Verification on Passive Structures

Before diving into the characterization of the active device, first let us focus on the behavior of the test structures that are used for transistor de-embedding. Fig. 4.9 displays the capacitances related to C-O (no-T0) (after classic and modified TRL calibration) and the analogous C-O (no-TM3) after a TRL calibration made by the dedicated M3 standards. We observe that the measurement follows the simulation's trend, both intrinsic and complete-probe. The "signature" of the PP-220 probe is present in all three plots in the 140-220 GHz range using the TRL calibration, leading to a negative C_{12} value.

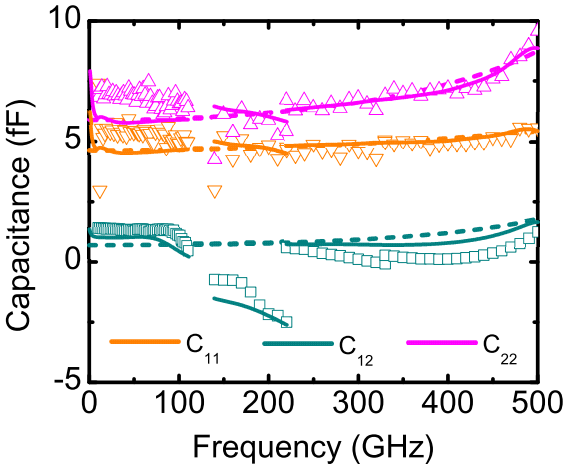
Comparing Fig. 4.9a to Fig. 4.9b, the port capacitance values decrease, since the portion of the line accounted for this parasitic is reduced; the reference plane position closer to the center of thru effectively removes the spurious inductive effect due to the access lines, that manifests itself as inconstant port and coupling capacitances over frequency. We notice that the complete-probe model of the capacitance value slightly differs from measurement, in the first frequency band, for the modified TRL as much as the classic TRL.

Now, Fig. 4.9b versus Fig. 4.9c let us compare the one-tier calibration to the classic approach. Port capacitances are constant and yield similar values to the modified TRL, just with higher noise. The fact that the values are so similar to the modified TRL confirms that in both cases they are mainly due to the metallic elements of the interconnect's design above the transistor's regions. In the case of C-O (no-T0) the final level of this interconnect, the M1 footprint, is present as small copper bits close to each other, much closer than the rest of the BEOL metal stack. As for C-O (no-TM3), the M8-to-M4 metal stack is completely absent, but the small copper bits are left on M3 and M2, as well as M1, probably motivating the additional capacitance at LF.

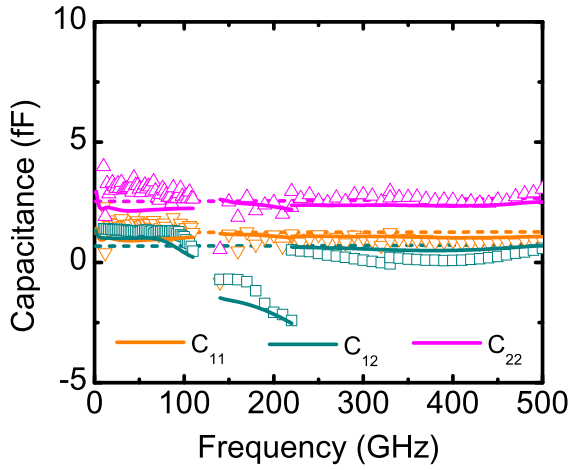
The complete-probe simulation captures very precisely the real calibrated trends, even at LF; however, C_{22} simulated curve stands out with an inaccurate offset. Both measured and simulated C_{12} are lower in the M3 case and slightly negative. Negative capacitance at HF shows the limits of TRL calibration and may affect de-embedded measurements (f_{max}, for instance), but since the actual purpose of the M3 structures is to avoid to perform de-embedding, this is just relevant within our comparative study.

Through Fig. 4.10 we are able to provide additional information on this layout. It is worth commenting that for these structures and these structures only, on both measurements and simulations, the reflects used are: pad-short instead of pad-open and short-M8 instead of open-M8. Less accurate and more noisy measurements, as well as non-physical simulated curves have been observed when using opens, like elsewhere.

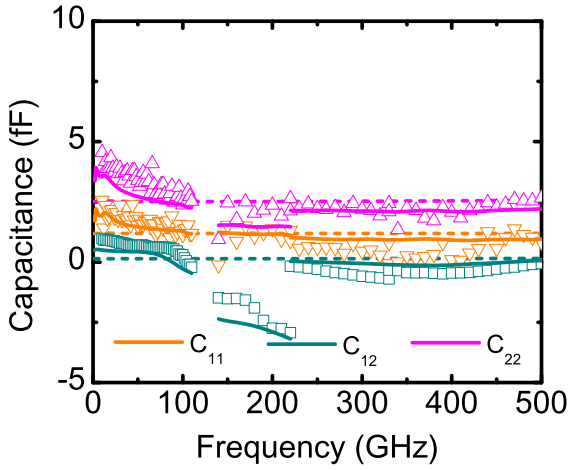
We first evaluate the inductances of C-S with classic (Fig. 4.10a) and modified TRL (Fig. 4.10b). Port inductances are 3 times higher than the case with S-M8 as a reflect, due to the M8 access lines contribution (from around 12-14 pH to 3-4 pH for port 1 and port 2); also, at HF, the classic case is frequency-dependent due to non-inductive parasitics. Ground inductance L_3 is almost unchanged.



(a) C-O (no-T0) after classic TRL (reflect: P-O) cal



(b) C-O (no-T0) after modified TRL (reflect: O-M8) cal



(c) C-O (no-TM3) after M3-TRL cal

Figure 4.9: All the capacitances of the lumped-element circuit model linked to C-O. Reference plane position: at pad-open's edge (a), at open-M8's edge (b), at pad-open (M3) edge (c). Symbols: measurement, dashed: intrinsic simulation, solid: complete-probe simulation.

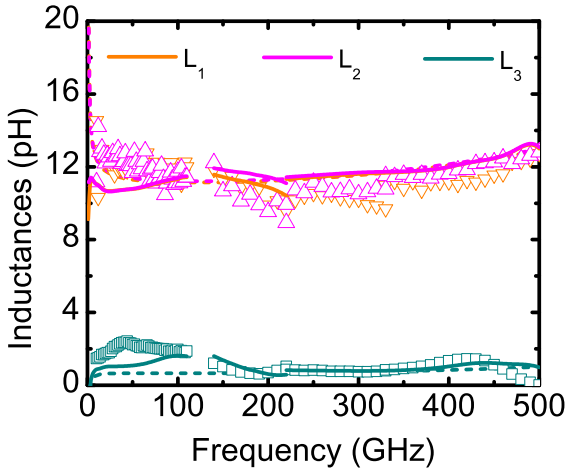
Comparing the modified TRL (Fig. 4.10b) and the M3 TRL (Fig. 4.10c), we observe that even though the intrinsic curves are only few femtohenry different, the measurement appears even better in terms of predictability, for all three inductances: the M3 layout gains in precision by removing the inductive path of the metal BEOL stack completely.

4.1.4 Calibration Verification on the Transistor

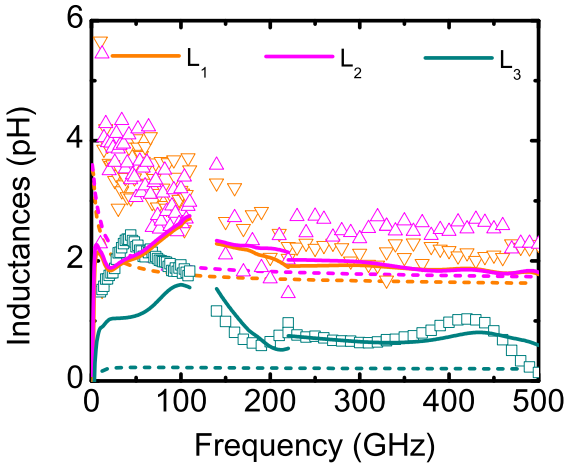
For a complete transistor analysis, we show below the measurements and HICUM simulation of the S-parameters and the two notorious figures of merit, f_T and f_{max} at a particular bias, i.e. V_{CB} = 0V, V_{BE} = 0.9V, given that similar observations can be made at the other operating points. Let us also show the transistor capacitances for a more complete point of view on the measurement performance with different vertical/horizontal reference plane positions.

The reader should be warned that, although the HICUM simulation is provided in all the following plots, its trend refers to the inner device performance. On the other hand, the measurements, which have been calibrated and de-embedded as we will describe more precisely in a moment, set the final reference to the M1 metal footprint of C-O (no-T0), instead of removing all the bottom metallization traces, like it was previously made by C-O (T-0).

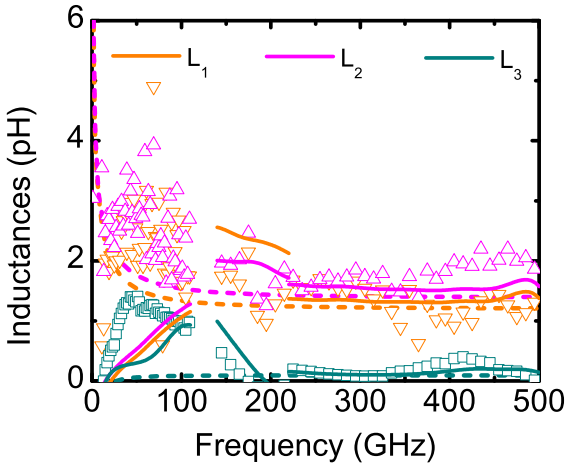
That is because an analogous C-O (T-M3) has not been implemented for de-embedding so far (and it would not be required either, provided that the M3 layout proves to be effective and



(a) C-S (T-0) after classic TRL (reflect: P-S) cal



(b) C-S (T-0) after modified TRL (reflect: S-M8) cal



(c) C-S (T-M3) after M3-TRL cal

Figure 4.10: All the inductances of the lumped-element circuit model linked to C-S. Symbols: measurement, dashed: intrinsic simulation, solid: complete-probe simulation.

behaves as expected). Yet a coherent comparison of the different two-/one-tier calibration would require identical setups to locate the reference plane in the exact same positions.

Finally, differences between the compact model simulation and measurements in the following will be possibly non-negligible, as already discussed with Fig. 3.15.

It is also important in this preamble to list what situations and position of the reference plane the following curves depict (see Table 4.2):

- "Classic": the DUT is the HBT with $A_E^{T0} = 0.2 \times 5\ \mu m^2$ and classic BEOL (T-0). The reference plane is located at the P-O metal edges after calibration, at M1 level after de-embedding. This will be considered as the reference;
- "Classic O-M8": the DUT is the HBT with $A_E^{T0} = 0.2 \times 5\ \mu m^2$ and classic BEOL (T-0). The reference plane is located at the O-M8 metal edges after calibration, at M1 level after de-embedding;
- "M3 (2-tier)": the DUT is the HBT with $A_E^{T0} = 0.2 \times 5\ \mu m^2$ and M3 BEOL (T-M3). The reference plane is located at the P-O (M3) metal edges after calibration, at M1 level after de-embedding;
- "M3 (1-tier)": the DUT is the HBT with $A_E^{T0} = 0.2 \times 5\ \mu m^2$ and M3 BEOL (T-M3). The reference plane is located at the P-O (M3) metal edges after calibration. The vertical position of the

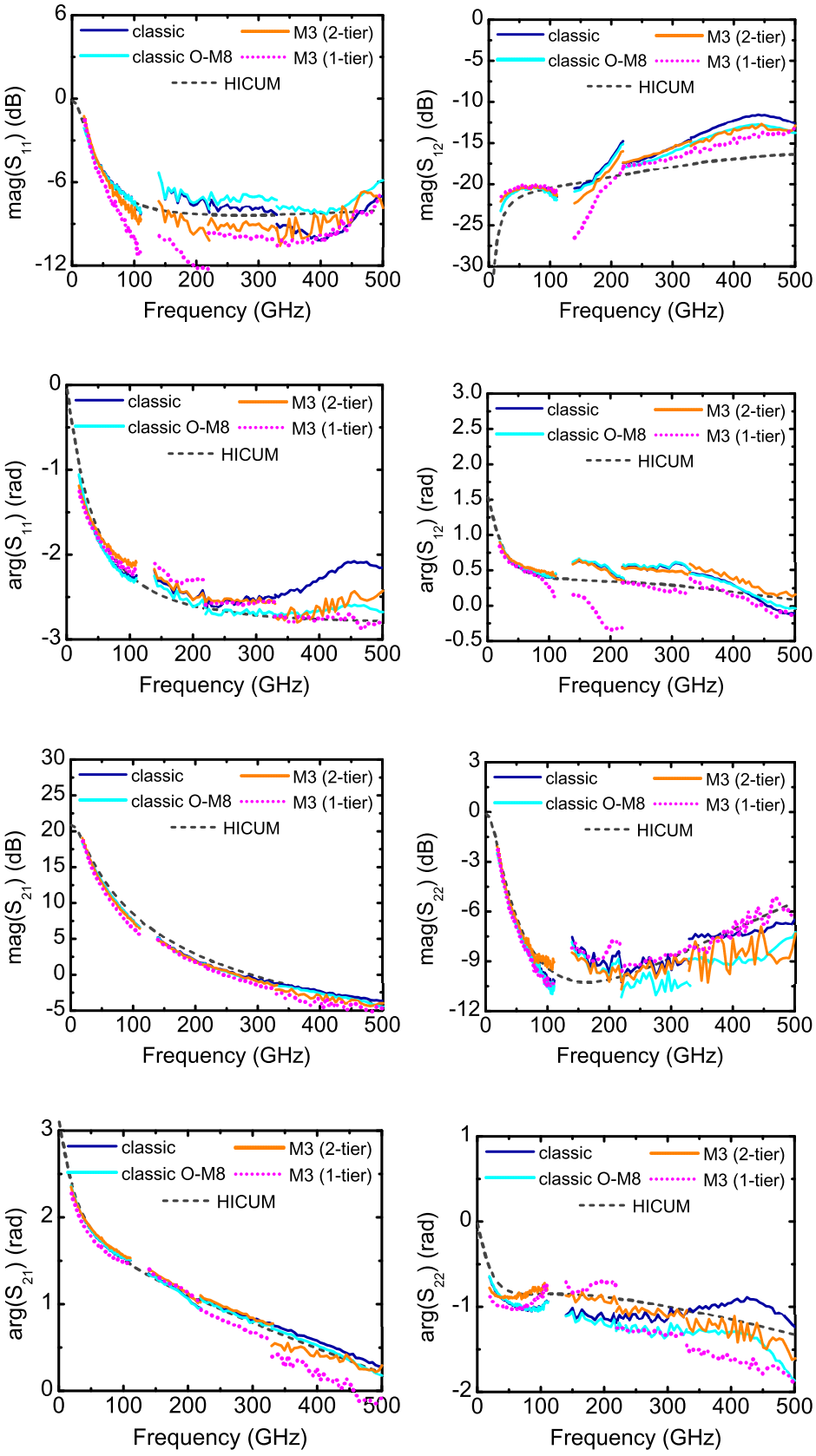


Figure 4.11: HICUM model compared to calibrated measured S-parameters of the HBT (V_{CB} = 0V, V_{BE} = 0.9V): classic TRL w/P-O + de-embedding w/C-O (no-T0) ("classic"); classic TRL w/O-M8 + de-embedding w/C-O (no-T0) ("classic O-M8"); M3 TRL + de-embedding ("M3 2-tier"); M3 TRL ("M3 1-tier").

Legend Name	HBT	Calibration			Z0 correction	De-embedding	
		Thru	Reflect	Line		Short	Open
Classic	T-0	Thru	P-O	L-110G/ L-500G	P-L	C-S (T-0)	C-O (no-T0)
Classic O-M8	T-0	Thru	O-M8	L-110G/ L-500G	P-L	C-S (T-0)	C-O (no-T0)
M3 (2-tier)	T-M3	Thru (M3)	P-O (M3)	L-110G (M3)/ L-500G (M3)	P-L (M3)	C-S (T-M3)	C-O (no-TM3)
M3 (1-tier)	T-M3	Thru (M3)	P-O (M3)	L-110G (M3)/ L-500G (M3)	P-L (M3)	-	-

Table 4.2: Description of used standards for different calibration approaches.

reference plane is therefore different from the other cases: this is why the corresponding curve is dotted in the following plots.

Let us start off with the S-parameters (Fig. 4.11). As for magnitude, the classic O-M8 curve perfectly matches in the lower part of the spectrum the classic case, and progressively distances itself in the third and fourth band particularly, yet it keeps quite constant all over frequency, varying its slope only after 440 GHz ($\mathrm{mag}(S_{11}), \mathrm{mag}(S_{12}), \arg(S_{11}), \arg(S_{12}), \arg(S_{22})$).

On the phase of S_{11} as well as S_{22}, the HF bump of the classic curve is almost completely suppressed by the O-M8 calibration. In one case (the classic) we de-embed a distributed open and short because the line access is not negligible, while in the other case, the reference plane is moved closer to DUT and de-emebedding structures. The open and short are constant and better captured by a lumped model.

As for the M3 layout, we see an unexpected offset between classic and M3 (2-tier) on $\mathrm{mag}(S_{11})$. While the reference plane is in the same position, the curves are not perfectly identical, and are detached from the other starting at LF; the other parameters' curves are quite akin, except at very HF. We also remark that the 2-tier curve is overall noisier in the last band, although the horizontal position after the first tier calibration guarantees in this case the absence of dips.

By comparing the latter curve with the M3 (1-tier) curve, we can see the effect of the removal of the de-embedding stage: the first most notable effect is that the measurement is consistently less noisy, thanks to avoided additional matrix manipulation of data. Particularly at HF (but especially from 140 to 330 GHz) this avoidance appears beneficial, with more continuous trend and good replication of both the HICUM and de-embedded curves.

We do not remark any bump on the phase of S_{11}, the magnitude of S_{12}, nor the phase of S_{22}, finally tracing them back to inaccurate interpretation by the lumped-element de-embedding model.

The improved continuity brought by the 1-tier approach is lost in the 330-500 GHz, as clearly visible on the phase of S_{21} and S_{22}: we will try to explain this behavior below.

The performance of the 1-tier case is undermined in the 140-220 GHz range, probably demonstrating that, with this critical set of probes (PP-220), any de-embedding has a positive effect, since it is the only way to remove part of the probe coupling.

The M3 (2-tier) and classic O-M8 cases, both extending the post-calibration reference plane position compared to the classic approach, perform overall quite similarly, with noise stronger in the former, as a consequence of de-embedding (hard-to-measure parasitics of C-S and C-O). Apart from small HF artefacts, the classic approach still proves to perform well, matching very consistently at LF to the model.

The transistor's cold capacitances are plotted on Fig. 4.12.

As expected, the highest difference compared to the HICUM curve comes from the collector capacitance, which when it is uncorrected by C-O (no-TM3), embeds the port-to-port coupling capacitance as well. Overall, M3 (1-tier) also displays an offset (1.5-2 fF on C_{be}, 3-4 fF on C_{cs}, and almost absent on C_{bc}), roughly corresponding to C-O (no-TM3)'s capacitance: these differences affect both f_T and f_{max}, as we will see below.

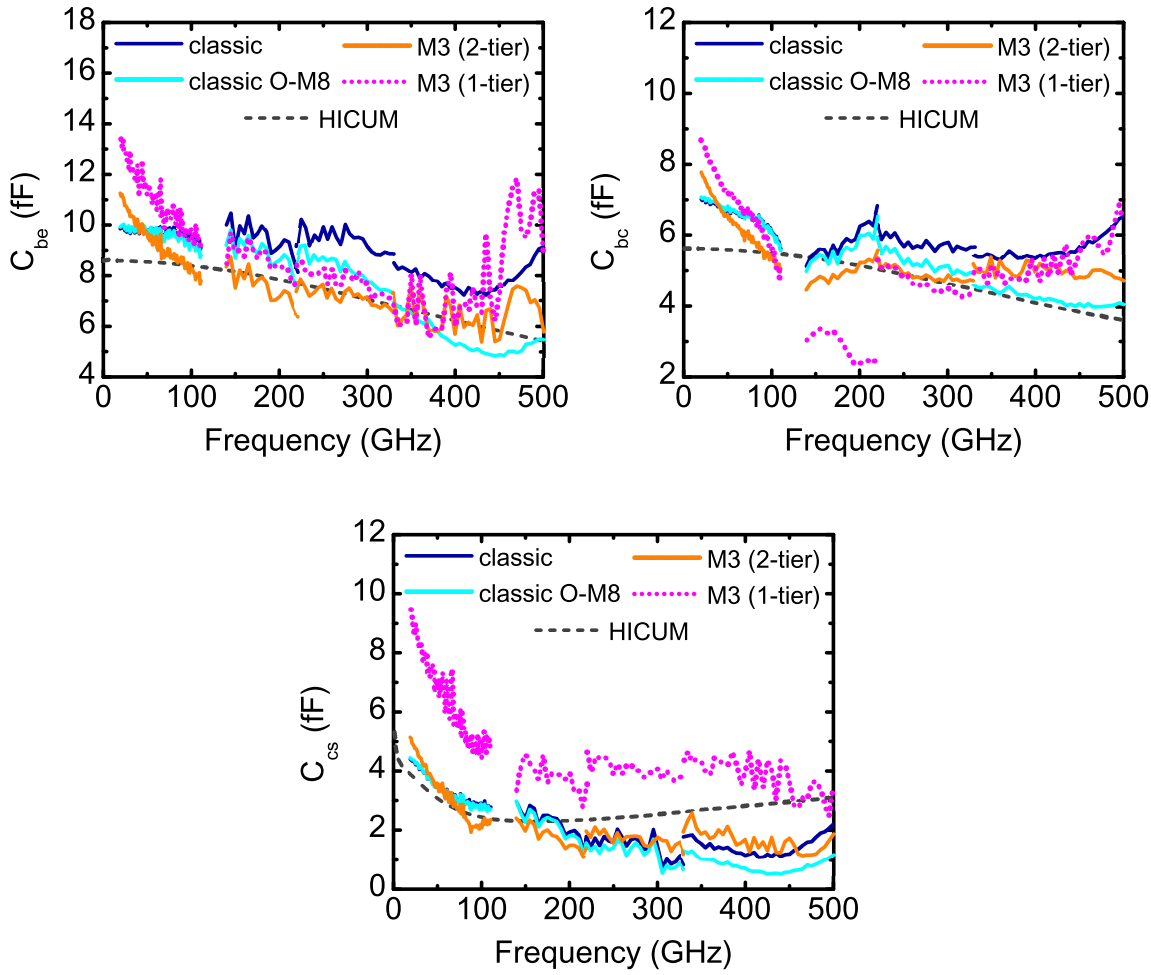


Figure 4.12: Transistor capacitances, comparing the HICUM model to calibrated measured data (V_{CB} = V_{BE} = 0V). Classic TRL w/P-O + de-embedding w/C-O (no-T0) ("classic"); classic TRL w/O-M8 + de-embedding w/C-O (no-T0) ("classic O-M8"); M3 TRL + de-embedding ("M3 2-tier"); M3 TRL ("M3 1-tier").

The high LF losses in the first band are clearly visible after both M3 calibrations, on every measured capacitance (similar frequency roll-off).

The M3 (1-tier) case shows high noise in the last band: this effect can be seen on the S-parameters but is magnified here. Moreover, the non-physical trend in the 140-220 GHz band, that we observed in many plots in Fig. 4.11, is here present on C_{bc} only, reinforcing the previous conclusion on coupling.

As for the two classic TRL calibrations, the curves start diverging increasingly from the second band, proving the effectiveness of a more extended first-tier calibration. Although less clear, the offset between the classic and M3 (2-tier) cases is still visible at least from approximately 60 to 350 GHz.

We conclude with the figures of merit (Fig. 4.13). Concerning the transit frequency, we can state that, in general, band continuity is preserved. An increase of f_T can be observed in the upper frequency bands due to the single-pole approximation.

The effect of the LF noise is not visible here. However, both the M3 cases are perturbed by high noise levels in the last band, as a result of the noisy capacitances in this same band, as we have just seen. This may be ultimately linked to inadequate power levels provided to our M3 lines in the last frequency band.

We examine finally the maximum oscillation frequency f_{max}. Let us consider the plot on Fig. 4.13. Here, the roll-off of f_{max} in the 1-110 GHz band highlights the increased line losses (we recall Fig. 3.34). From 140 to 330 GHz, however, f_{max} remains quite constant, with both classic and M3

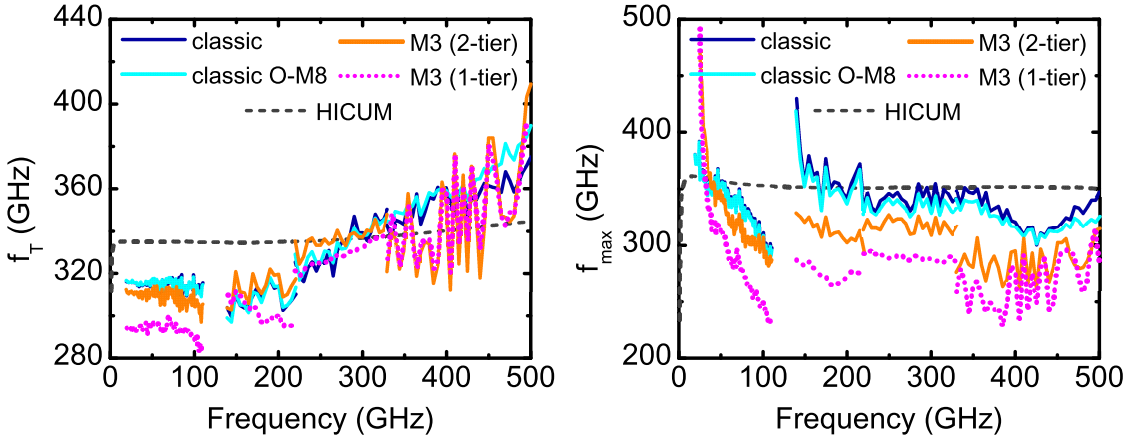


Figure 4.13: HICUM model compared to calibrated measured transit frequency and maximum oscillation frequency (V_{CB} = 0V, V_{BE} = 0.9V): classic TRL w/P-O + de-embedding w/C-O (no-T0) ("classic"); classic TRL w/O-M8 + de-embedding w/C-O (no-T0) ("classic O-M8"); M3 TRL + de-embedding ("M3 2-tier"); M3 TRL ("M3 1-tier").

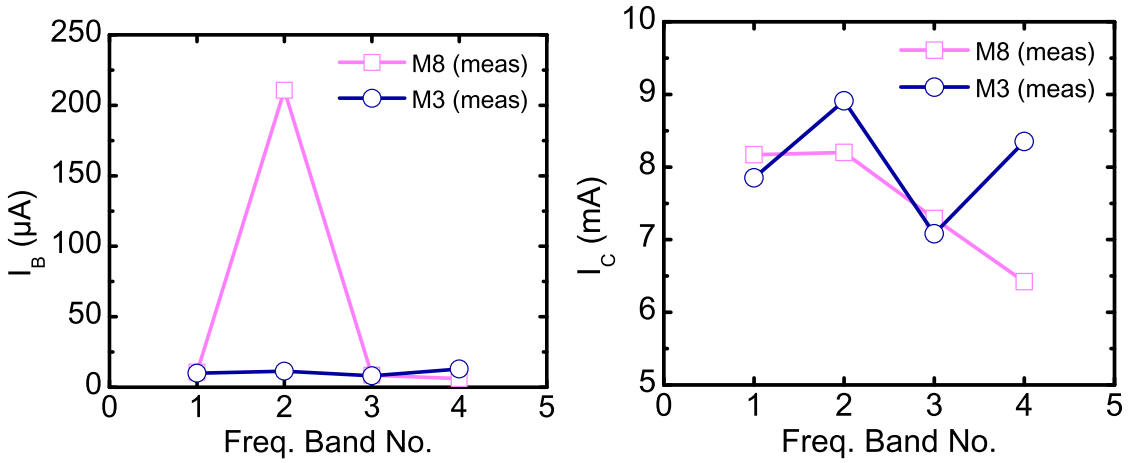
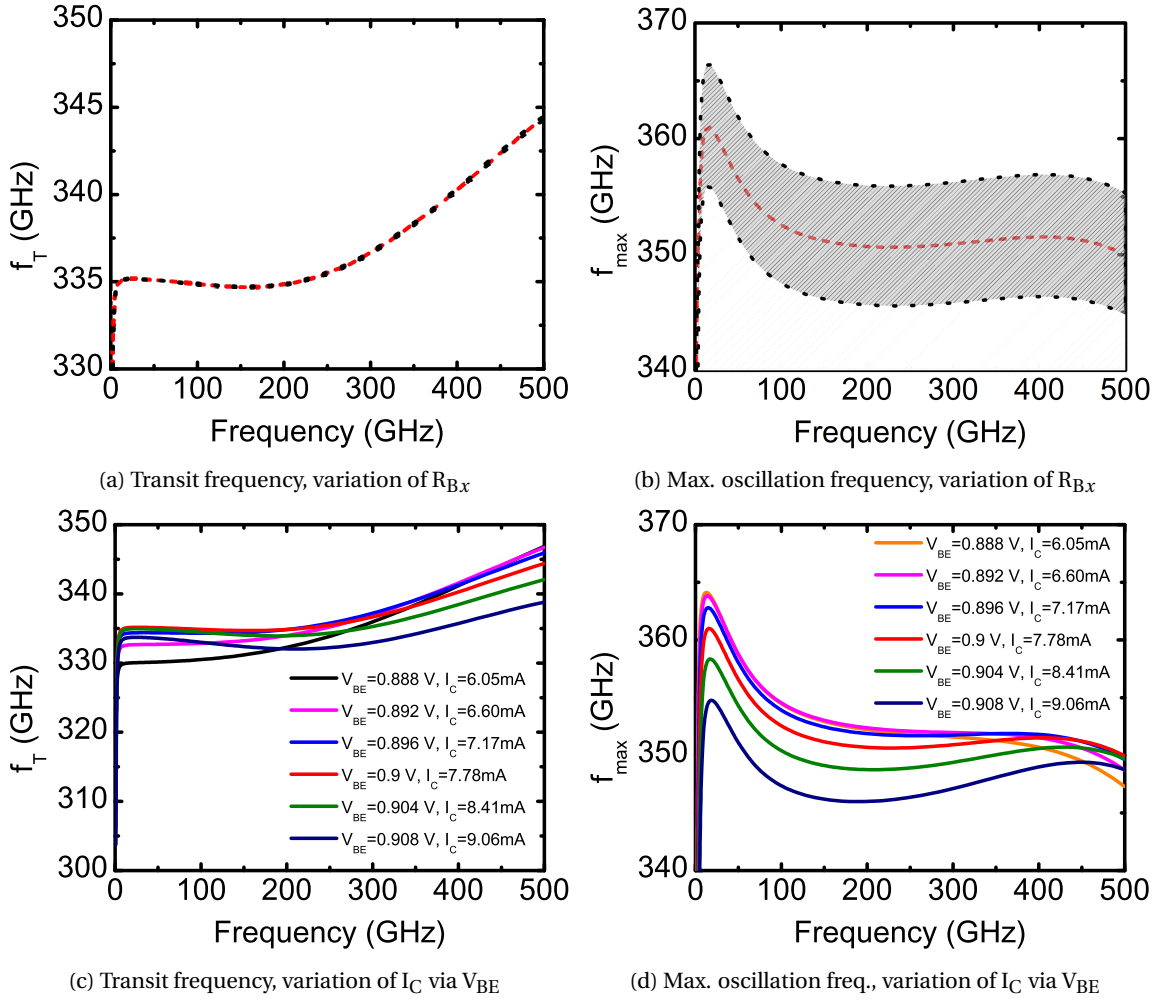


Figure 4.14: Base and collector DC current values flowing into the device during measurement on each frequency band: band 1 = 1-110 GHz; band 2 = 140-220 GHz; band 3 = 220-325 GHz; band 4 = 330-500 GHz. V_{CB} = 0V, V_{BE} = 0.9V.

TRL. Nevertheless, we observe again the offset between the classic TRL and the M3 TRL with de-embedding. As seen, f_{max} is very sensitive to impedance correction, especially the imaginary part of Z_0, and to contact quality and bias, as we will demonstrate in a moment. Multiple, optimal quality measurements would be required to verify this trend.

Let us take a look to the DC operating point (Fig. 4.14) for better understanding the differences we observe. First, we might be tempted to link the much higher base current measured for the M8 case to some variation on the HF parameters. The anomalous current value in the 140-220 GHz band is the symptom of an internal degradation of the transistor as a result of excessive stimulation, possibly a deterioration the base junction. Anyway, this does not prevent the transistor from working.

We believe, however, that there is a correlation between the average value of f_T and f_{max} and the base external resistance R_{Bx}: if any contact changes during calibration or de-embedding and the transistor measurement, due to lack of repeatability, a delta of resistance will be transferred to the transistor. We can imagine to model this by a delta variation of the extrinsic base resistance R_{Bx}. As observed on the DC measurements of pad-short, the uncertainty of the total measured DC resistance may vary up to 0.5 Ω. In Fig. 4.15a and 4.15b we show what is expected by the HICUM model for a variation of $\Delta R_{Bx} = \pm 0.5\,\Omega$: while f_T is unaffected, Δf_{max} is up to 10 GHz.



(a) Transit frequency, variation of R_{Bx}

(b) Max. oscillation frequency, variation of R_{Bx}

(c) Transit frequency, variation of I_C via V_{BE}

(d) Max. oscillation freq., variation of I_C via V_{BE}

Figure 4.15: FoM from HICUM after variation of the external base resistance and collector current. The reference curve is shown in red.

On the right panel of Fig. 4.14, we see the collector current. We recall that the collector current is correlated to the trend of the figures of merit. Let us show Fig. 4.15c and 4.15d. Here, we have modified the input base voltage so to provoke a variation of the output collector current as experienced in our measurements (in the range between 6 and 9 mA). Lower current variation affect f_T mainly at LF, while higher current variation at HF, approximately by the same quantity (10 GHz). Variations by higher currents, on the other hand, affect f_{max} mainly at LF, by 10 GHz at the most. Although it is not trivial to correlate exactly each contribution to a specific effects, these observations have to be accounted for measurement quality improvements.

Going back to Fig. 4.13, we observe coherent results with both classic approaches, with few differences when displacing the reference plane (O-M8). The dip in the last band of f_{max}, although it doesn't make the overall curve discontinuous, is not physical. We link it to the wear on the pads of the L-500G, and we recall that this artefact was also visible on the attenuation constant (compare with Fig. 4.4). RF pads covered by or made of gold are proved to provide better contacts.

Concerning the M3 (1-tier) case, we observe on both f_T and f_{max} yet another offset. The initial suggestion was that the influence of the BEOL up to M3 was negligible, since it exhibits very small values of C and L, but it turns out that these small parasitic values cannot be completely neglected and de-embedding must be performed anyway, under penalty of loosing up to 50 GHz of f_{max}.

Yet, if we observe the M3 TRL curve without de-embedding, its trace is generally flatter and better outlined, reinforcing our original motivation of a one-tier calibration in order to avoid extra matrix computation, a source of possible additional errors and noise.

4.1.5 One-Tier Calibration at M1, an Overview: 3D TRL

A tentative of one-tier calibration where the reference plane is brought to transistor level is accomplished by the 3D TRL, first introduced by Potéreau *et al.* in [81]. The name of this TRL approach stems from the topology of the designed lines, which are the only calibration standards changing from the classic TRL (pad-open, pad-load and DUT's BEOL are the same). The lines propagate the signal in part vertically (i.e. along the z-axis) with respect to the ground plane, during the descent to M1.

In Fig. 3.9 and Table 3.2 they are located at position C7 (thru 3D), A10 (L-110G 3D), A3 (L-500G 3D). As presented in [81], this approach allows to remove the de-embedding phase and provides good performance compared to standard TRL associated with complex de-embedding (it has been shown up to 67 GHz).

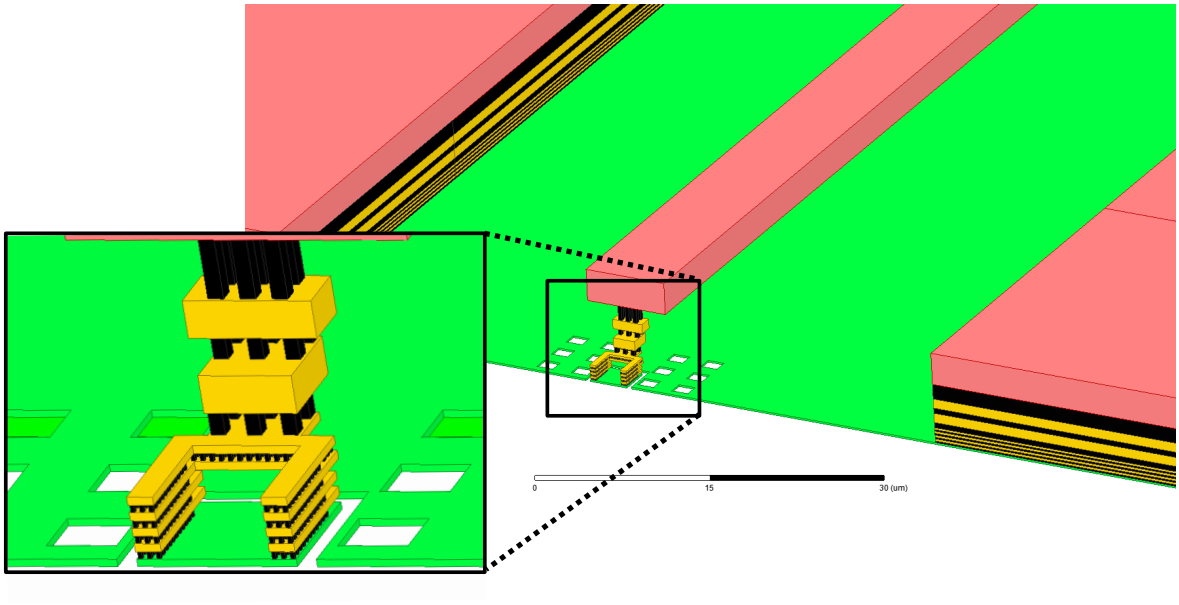


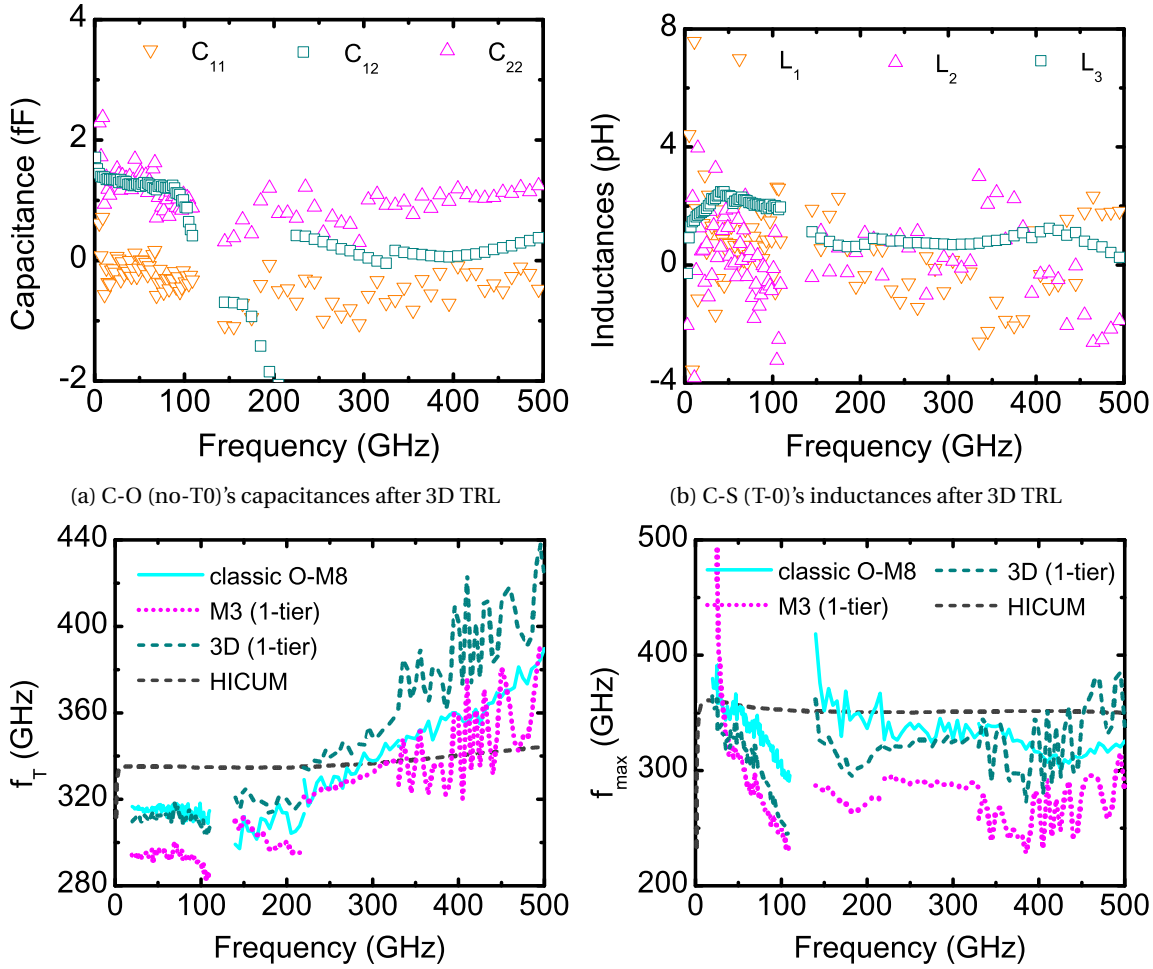
Figure 4.16: Half HFSS model of run 2's L-500G (3D). The model is mirrored to the reader's direction, not shown for clarity. Color key: pink is M8 (copper), lime indicates the M1 ground (copper), yellow are layers M2-M7. The plate located at M1 below the metal connections, joins port 1 BEOL to port 2 BEOL. Dielectric layers are not shown.

Let us first introduce the structure of the lines. Fig. 4.16 shows half of L-500G line. The signal trace (in pink), once reached the center where the transistor would lay, descends to M1 with the same BEOL stack as the transistor T-0 and C-O/C-S. It reaches M1, where a small metal plate connects the base and collector terminals, providing a transmission path to port 2. The (almost) symmetrical half line to port 2 is not displayed in the picture, for clarity. In essence, the C-O topology has been taken, a plate at M1 is added, and access lines are stretched to match the true line lengths.

It is true that such a geometry violates the microstrip topology, thus no TEM field propagates in its entirety. Moreover, the line is also not symmetrical nor homogeneous, and the energy partially flows on less dense metallic connections (vias).

Even though we concede many compromises with such a structure, the violation of the microstrip line hypothesis is confined to only approximately 9% of its total length. By intrinsic simulation we did not find any major deviation on total losses (not even G), nor propagation constant, nor higher mode generation, compared to classic lines.

Hence, we decide to use these three lines to calibrate the usual set of raw measurements and compare it to the classic approach (with reflect = O-M8, due to its proven good performance) and one-tier calibration with M3 TRL. The HBT under test is once again T-0, and to stick to the comparison, we will use O-M8 as a reflect for the 3D TRL as well.



(a) C-O (no-T0)'s capacitances after 3D TRL

(b) C-S (T-0)'s inductances after 3D TRL

(c) T-0's transit frequency

(d) T-0's max. oscillation freq.

Figure 4.17: Use of 3D TRL compared to previous approaches for calibrating measurements: classic TRL w/O-M8 + de-embedding w/C-O (no-T0) ("classic O-M8"); M3 TRL ("M3 1-tier"); classic TRL w/3D lines ("3D 1-tier"). The non-univocal position of the reference plane is indicated with different types of hatching.

Fig. 4.17 presents C and L of C-O (no-T0) and C-S, respectively, together with the figures of merit of the transistor at peak-f_T. It is hard to define the values of the intrinsic capacitances /inductances, since the reference plane is theoretically located exactly in the middle of the thru. We expect the port inductances and capacitances to be null (while C_{12} and L_3, not necessarily).

However, due to loss of symmetry of our structures, compared to the assumptions of a TRL calibration, we observe discrepancies from the theoretical zero value. In fact, while we remark that C_{11} is slightly negative (-0.5 fF on average) but overall close to zero, C_{22} is comprised between 0 and 2 fF, essentially because of the large footprint of M1 collector contact. Also, C_{12} is essentially unchanged compared to the previous analysis.

As for the inductances of C-S, the average of L_1 and L_2 is zero, even though they are noisier than the previous cases. The ground inductance is unchanged.

On the figures of merit, we see an improvement compared to the M3 TRL with one-tier on both f_T and f_{max}, making this approach a promising alternative for one-tier calibration.

We notice that at HF values, f_T's 3D TRL result is even higher than the classic TRL, thanks perhaps to overall low values of transistor capacitances after calibration.

So, let us look at Fig. 4.18 for a comprehensive evaluation of their trends. All three capacitances are globally lower than or close to the two-tier procedure, except C_{bc}, since the TRL calibration alone does not correct crosstalk. In [81], it has been proposed to add a de-embedding step for the correction of this particular parasitic contribution.

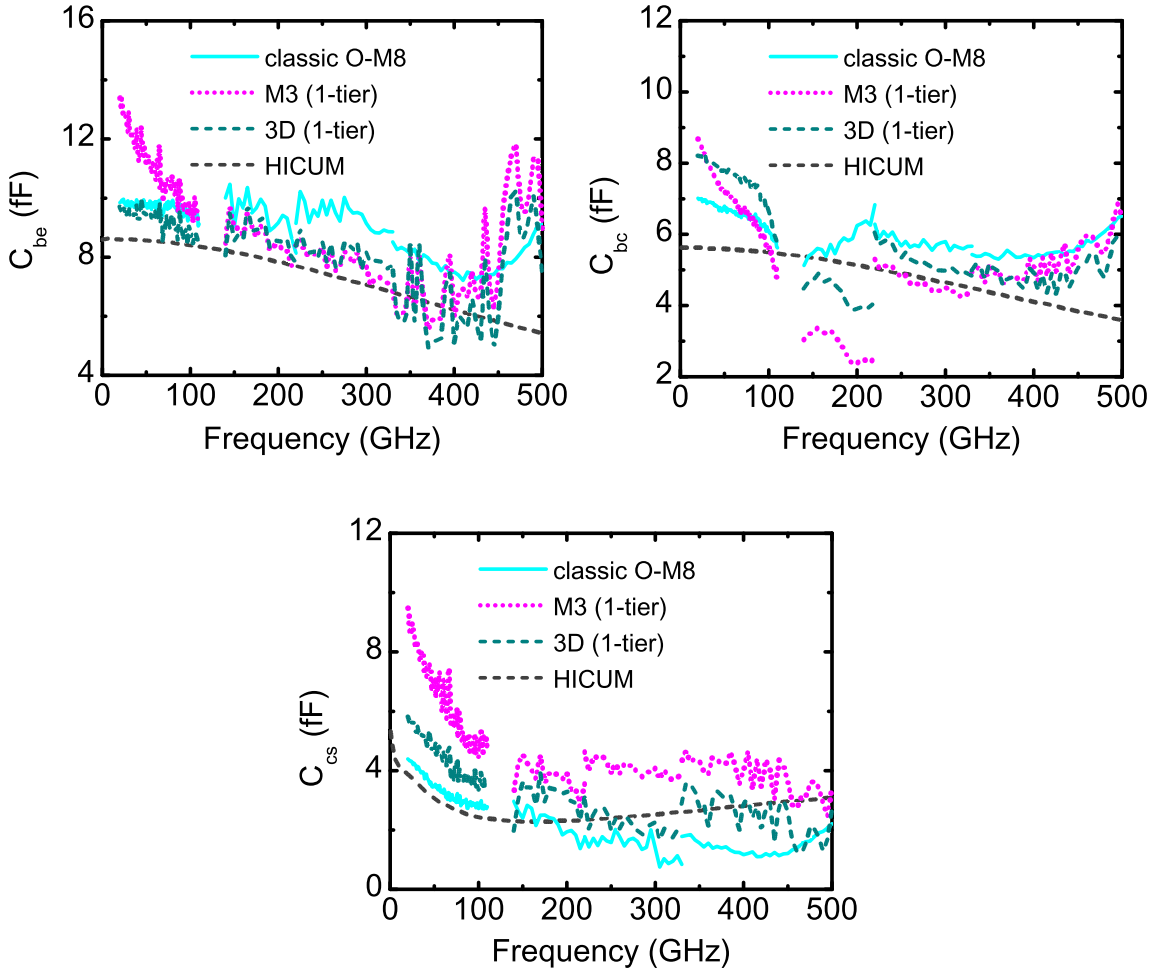


Figure 4.18: Use of 3D TRL compared to previous approaches for calibrating measurements on transistor capacitances ($V_{CB} = V_{BE} = 0$V): classic TRL w/O-M8 + de-embedding w/C-O (no-T0) ("classic O-M8"); M3 TRL ("M3 1-tier"); classic TRL w/3D lines ("3D 1-tier"). The non-univocal position of the reference plane is indicated with different types of hatching.

The 3D TRL proves to be a valuable choice for one-tier calibration, and further research for optimising the lines may definitely make it a strong candidate for innovative TRL calibration technique at millimeter wave frequencies.

4.2 Lines with Constant Inter-Probe Distance: the Meander Layout

We have extensively shown that the TRL calibration is best suited, for several reasons, to deliver accurate millimeter wave measurements. This method, however, relies on measuring at least two line standards (the line and the thru, but necessarily one or two other to span to lower frequencies, not to mention mTRL, which uses up to 5-6 lines) which have their own lengths. The operator is thus forced to accommodate the RF probes on a new pad position. The calibration repeatability is therefore degraded and the error terms calculation might be compromised, thus leading to discrepancies and questionable conclusions on measurements.

Potereau *et al.* [62] set up an experiment to assess the impact of axial probe displacement and observed a 50% variation of the admittance on the port of the displaced probe and a related capacitance drop. They concluded that due to reduced probes' crosstalk, different cable position and other small modifications of the measurement environment, one is exposed to less accurate HF measurements.

Recent studies have been led with a focus on contact repeatability and discussions on mea-

surement uncertainty of on-wafer measurement and probe positioning [101, 79]. Additionally, a fully-automated millimeter-wave measuring system may work better if all the structures are designed by keeping a constant inter-probe distance. Finally, referring back to Table 3.3, we observe that the L-500G and L-110G that we use for the run 2 TRL calibration consume 1.85 and 8.15 times more area than the other test structures on the wafer, respectively, and due to their different geometry, they do not allow a fully chessboard-like layout.

Therefore, a novel architecture has been conceived, in which transmission lines are designed as meanders [151], i.e. the portion of the upper metal layer carrying the signal from port-1 to port-2 (and vice-versa) is rolled up perpendicularly to probe axes using a unit pattern of 45 degrees angle transmission line, as introduced in [82]. This allows to design lines longer than the thru, as prescribed by the TRL algorithm, and keep the same inter-probe distance.

Compared to [82], the following work extends the method up to 500 GHz and rigorously highlights the meander lines' characteristics. We will also employ the usual approach combining measurements, EM simulation and co-simulation to interpret them.

4.2.1 The Meander Lines

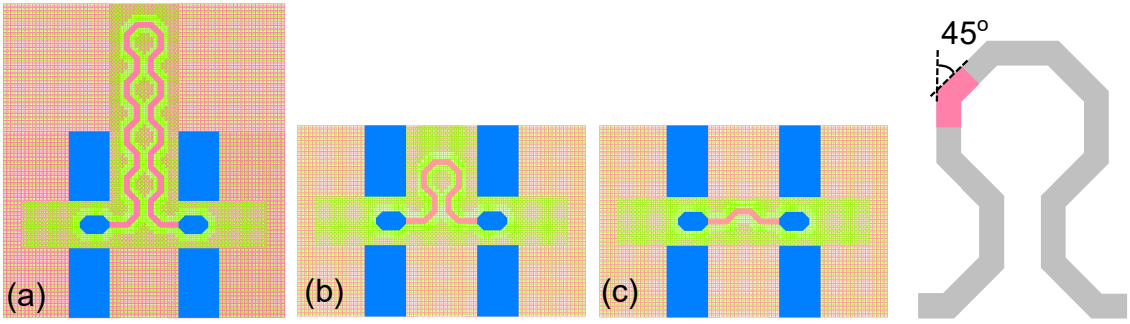


Figure 4.19: Top view of the on-wafer meander TRL calibration kit. L-110G (M) (a), L-500G (M) (b), thru (M) (c). L-500G's signal trace is shown on the right, with one of the "unit cells" highlighted.

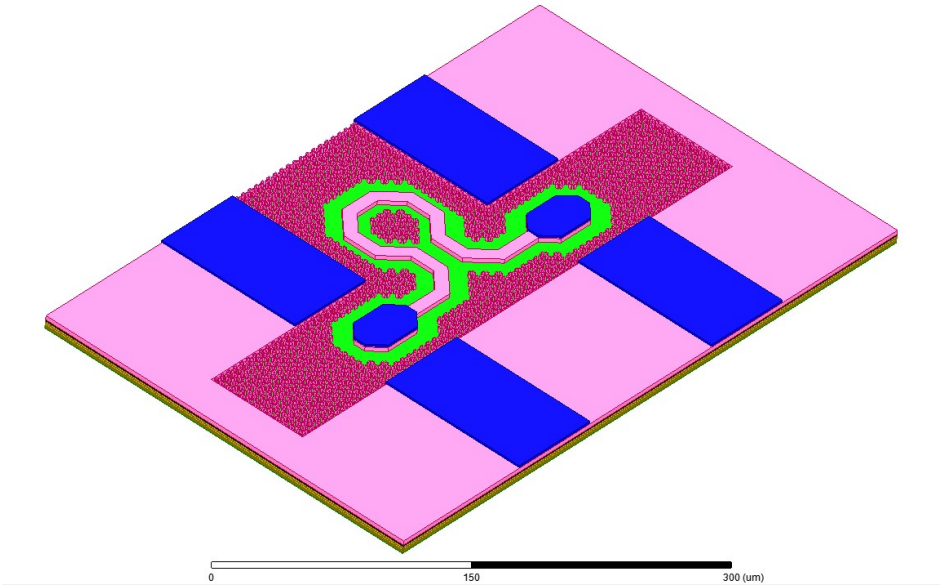


Figure 4.20: Artwork of L-500G. Color key: pink is M8 with dummies (copper), blue are pads (aluminum), lime indicates the ground (copper). Silicon dioxide is not shown for clarity.

Let us introduce the meander line calibration kit in Fig. 4.19. The location of the structures in Fig. 3.9 (Table 3.2) are C5 (thru), C11 (L-110G) and B6 (L-500G).

	Unit Cell	Thru	L-110G	L-500G
No. of cells	1	4	36	12
Shortest Path (µm)	15.84	63.36	570.24	190.08
Middle Path (µm)	19.04	76.16	685.44	228.48
Longest Path (µm)	22.24	88.96	800.64	266.88

Table 4.3: Actual lengths of meander elements.

As originally described in [82], the aim of the meander-type design is to avoid pushing back and forth the probes when measuring a standard with a different probe-to-probe distance, such as the straight line L-110G and L-500G. As a matter of fact, we recall that the port 1 to port 2 distances of these structures are 595 µm and 185 µm.

To perform a TRL calibration, one is confronted with the need of manually withdrawing the RF measurement probes and lay them back down to the pads, since the port-to-port distance of every other structure in our wafer is 65 µm. Therefore, lines creating a meander-type path for the signals have been designed.

This is done in order to maintain the inter-probe distance and avoid displacing the probes. As a useful consequence, and the on-wafer calibration standards' occupation is reduced (and in turn, the production costs), and no calibration imperfection due to mechanical movement in RF cables and mmW heads is introduced. Moreover, the impact of X or Y probe misplacement degrades the measurement accuracy, and can be mapped to a phase error [9] and propagated to calibrated data; by keeping the spacing constant we expect to reduce measurement uncertainty. However, such a benchmarking goes beyond the purpose of this presentation.

Each line (located at top metal level, M8) is drawn by pulling together multiple times a single element called "unit cell". The unit cell (highlighted in rose on the left of Fig. 4.19) consists of a piece of copper track routed at a 45 degree angle (a mandatory design rule at microwave frequencies) and assembled to form sufficiently long lines presenting quasi-octagonal patterns (see Fig. 4.20).

While in a classic straight line approach the lengths are univocally defined, we can determine multiple lengths for the cell (and consequently the lines), by considering either the shortest, the central or the longest path, due to the trace width. Table 4.3 lists the measured actual lengths of the lines.

Although we can state with certainty that the shortest path is linked to the lower boundary time after which the signal generated at one port reaches the other by flowing through the meander trace, we do not know how long we can consider each line for the TRL algorithm.

Nevertheless, since the phase constant β of a transmitted signal is only dependent on material properties and frequency, the signal must flow with different velocity (vector) but same speed (magnitude) on both the straight and the meander line, provided that both are conceived to work in the same frequency range. This observation translates into:

$$\beta_S = \beta_M = \omega\sqrt{LC} = \omega\sqrt{\mu\varepsilon} = \frac{\omega\sqrt{\varepsilon_{R,eff}}}{c_0} = \frac{\omega}{v_S} = \frac{\omega}{v_M} \tag{4.3}$$

where the "$_{S,M}$" subscripts indicate the type of line (straight, meander), and v is the phase speed of the transmitted signal. The previous equations are valid under the hypothesis of a TEM wave and low loss line, which are verified here in first approximation.

We recall Eq. 3.1, that alongside Eq. 4.3 yields:

$$l_{M,eff} = \frac{\phi_M}{\phi_S} l_S \tag{4.4}$$

where $\phi = \arg(S_{12}) = \arg(S_{21})$ is the phase shift of a line, linearly frequency-dependent.

$l_{M,eff}$ is what we call effective meander line length, and corresponds to the effective time delay for the signal to be transmitted from port to port on a meander line, with the same speed of the straight line case.

	Straight L. (μm)	Meander Effective L. (μm)					
		Intrinsic Simu.	VNA TRL	Case 1	Case 2	Case 3	Case 4
Thru	65	73	72	70	73	70	76
L-110G	595	620	660	610	613	607	613
L-500G	185	199	205	201	204	189	195

Table 4.4: Lengths of thru and lines.

Unfortunately, the lines' raw measured data do not provide any direct information on the value of ϕ, and the previous formula can be applied only for an indirect calculation. For example, one may want to use the intrinsic simulation to find the transmission parameters and evaluate the ratio.

Alternatively, we can use the VNA-calibrated data, namely the on-wafer TRL-calibrated results as they are computed by the algorithm inside the calibration software of the VNA. In both cases, an interpolation is applied to fit the curves from 1 to 500 GHz, compute their slopes and retrieve the effective meander length from the related straight line's length. The values we have found are shown in Table 4.4.

Needing either a simulator or the VNA data is impractical and potentially imprecise. For this reason, several calculations for direct extraction are presented below. The TRL calibration algorithm in the case of a non-zero thru, as we know, allows a direct calculation of $\gamma\Delta l$, i.e. the product of the propagation constant of two line with identical properties and the difference of their lengths, via the matrix [M] (refer to Eq. B.21 in Appendix B).

The raw data of a line and a thru can be manipulated to find the corresponding $\gamma\Delta l$, and by using $\overline{\beta}$, i.e. the linear interpolation of β found as described in Chapter 3, we can retrieve the line length difference, since the straight line one is known.

The pair of lines to begin the procedure with must be chosen so that one is a meander and the other a straight line. So if we consider just the cases where the shortest of the two lines is the thru, we are left with the following choices: (L-110G; thru (M)), (L-110G (M); thru), (L-500G; thru (M)), (L-500G (M); thru).

We exclude the cases employing the 500 GHz lines, since the interpolation will be performed in the same range as the interpolation for finding $\overline{\beta}$ (i.e. 28-56 GHz), and those lines would be outside of their range of validity.

For each of the remaining cases, two sub-cases are identified, depending on how we proceed extracting the effective length of the remaining lines. We consider a total of 4 cases:

- **Case 1**. From [M], we take the imaginary part of $\gamma\Delta l$, i.e. $\beta\Delta l = \beta\,(l_S^{110} - l_M^T)$, where l_S^{110} is the (known) length of L-110G and l_M^T is the (unknown) length of thru (M). We divide now by the extracted $\overline{\beta}$ from the classic TRL:

$$\frac{\beta\Delta l}{\overline{\beta}} = \Delta l$$

and we easily find the unknown term:

$$l_M^T = l_S^{110} - \Delta l$$

For the two other terms we follow the same procedure, entirely applied on the meander lines:

$$\mathrm{Im}(\gamma\Delta l) = \beta\,(l_{\mathrm{M}}^{110} - l_{\mathrm{M}}^{\mathrm{T}}) \Longrightarrow \frac{\beta\Delta l}{\bar{\beta}} = \Delta l$$

by which we find L-110G (M)'s length:

$$l_{\mathrm{M}}^{110} = l_{\mathrm{M}}^{\mathrm{T}} + \Delta l$$

and analogously:

$$\mathrm{Im}(\gamma\Delta l) = \beta\,(l_{\mathrm{M}}^{500} - l_{\mathrm{M}}^{\mathrm{T}}) \Longrightarrow \frac{\beta\Delta l}{\bar{\beta}} = \Delta l$$

to find L-500G (M)'s length:

$$l_{\mathrm{M}}^{500} = l_{\mathrm{M}}^{\mathrm{T}} + \Delta l$$

- **Case 2**. The same procedure is followed, by considering L-110G (M) and the straight thru as starting point:

$$\mathrm{Im}(\gamma\Delta l) = \beta\,(l_{\mathrm{M}}^{110} - l_{\mathrm{S}}^{\mathrm{T}}) \Longrightarrow \frac{\beta\Delta l}{\bar{\beta}} = \Delta l$$

to find L-110G (M)'s length:

$$l_{\mathrm{M}}^{110} = l_{\mathrm{S}}^{\mathrm{T}} + \Delta l$$

and from that, thru (M) and L-500G (M), exactly as before.

- **Case 3**. We proceed as in case 1 until we find $l_{\mathrm{M}}^{\mathrm{T}}$. Now, we recall the hypothesis under which all the cases are valid, namely $\beta_{\mathrm{S}} = \beta_{\mathrm{M}}$, from Eq. 4.3. The following equation is therefore true:

$$\Delta l_{\mathrm{M}} = \frac{\beta_{\mathrm{M}}\Delta l_{\mathrm{M}}}{\beta_{\mathrm{S}}\Delta l_{\mathrm{S}}}\,\Delta l_{\mathrm{S}}$$

The numerator and the denominator of the fraction can be found and Δl_{S} is well known in any case. So if we first use (L-110G (M); thru (M)) we can find the term at the numerator and by (L-110G; thru), the term at the denominator. Eventually, $l_{\mathrm{M}}^{110} = l_{\mathrm{M}}^{\mathrm{T}} + \Delta l_{\mathrm{M}}^{110}$. The same is valid with (L-500G (M); thru (M)) for the numerator and (L-500G; thru) for the denominator. Hence, $l_{\mathrm{M}}^{500} = l_{\mathrm{M}}^{\mathrm{T}} + \Delta l_{\mathrm{M}}^{500}$.

- **Case 4**. In this case, we find $l_{\mathrm{M}}^{\mathrm{T}}$ as in case 2 and we continue as in case 3.

The extracted values of lengths are all reported in Table 4.4. To determine which one of these values is best suited to use in our TRL calibration toolkit, we tested them to calibrate some actual measured raw data from run 2 layout (see Fig. 4.21).

With the exception of the first frequency range, where case 2 outperforms, case 3 works more appropriately overall, hence we choose it for the following meander line calibrations.

We observe that the impact of choosing one method over another only affects the usually very noisy C-S L_1 measurement by less than 20% and the HBT's figures of merit by even less than 0.5%. The difference in value on the inductance is due to the shift of the reference plane according to the lengths of the line considered. But since it's the same reference plane for the transistor, it actually has no impact on f_T and f_{max}: any deviation is corrected by de-embedding.

We conclude that, to minimize the error of the solely calibrated structure in the whole spectrum, case 3, 2 or the TRL VNA case should be preferred. In the following, case 3 will be used.

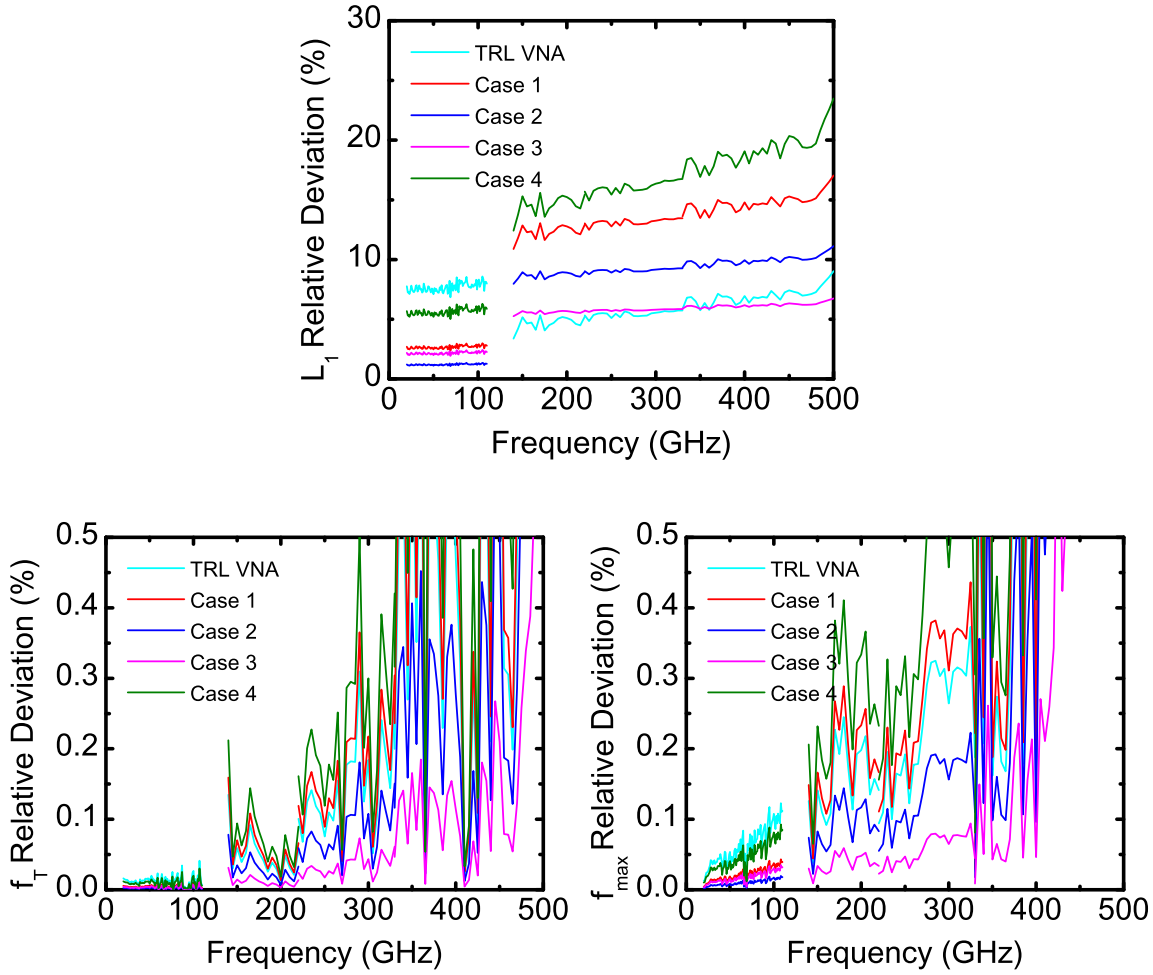


Figure 4.21: Different approaches to compute the effective meander lines' lengths: relative deviation w.r.t. the case of lengths computed by the intrinsic simulation.

We look at the electrical characteristics of the shortest (straight and meander) lines in Fig. 4.22 up to 500 GHz. All these parameters are consistent with the straight line case. Overall, the measurements in the first band are less noisy, probably hinting to a better probe contact or less worn-out pads.

The difficulties on measuring α, Z_0 and C at HF are the same for the meander line (noise and oscillations). While on the straight line the attenuation constant begins to decrease at around 350 GHz, becomes non-physical and finally rises again from 460 GHz on, the meander attenuation behaves similarly, though just increasing from 420 GHz on.

The interpretation we have given in the case of the straight line was associated to the wear of the pads, and this may indeed explain why the blue curve in figure falls below 0 dB. However, we are witnessing to an increase of the losses in both cases. Below, we will try to provide another explanation for these trends.

4.2.2 De-embedding Standards and Transistor's Characterization

Fig. 4.23 shows the measured and simulated capacitances of C-O (T-0) after TRL calibration in which classic straight and meander-type designs are compared. Note that C-O is physically the same for both straight and meander cases. As previously mentioned, what is changing are just two out of four TRL calibration standards, i.e. thru and lines; since the reflect and pad-load (for impedance correction) are identical.

Simulations are derived from both probe and intrinsic EM approaches (solid and dashed lines, respectively). Both measured curves show quasi-identical trends, slightly diverging in the last two

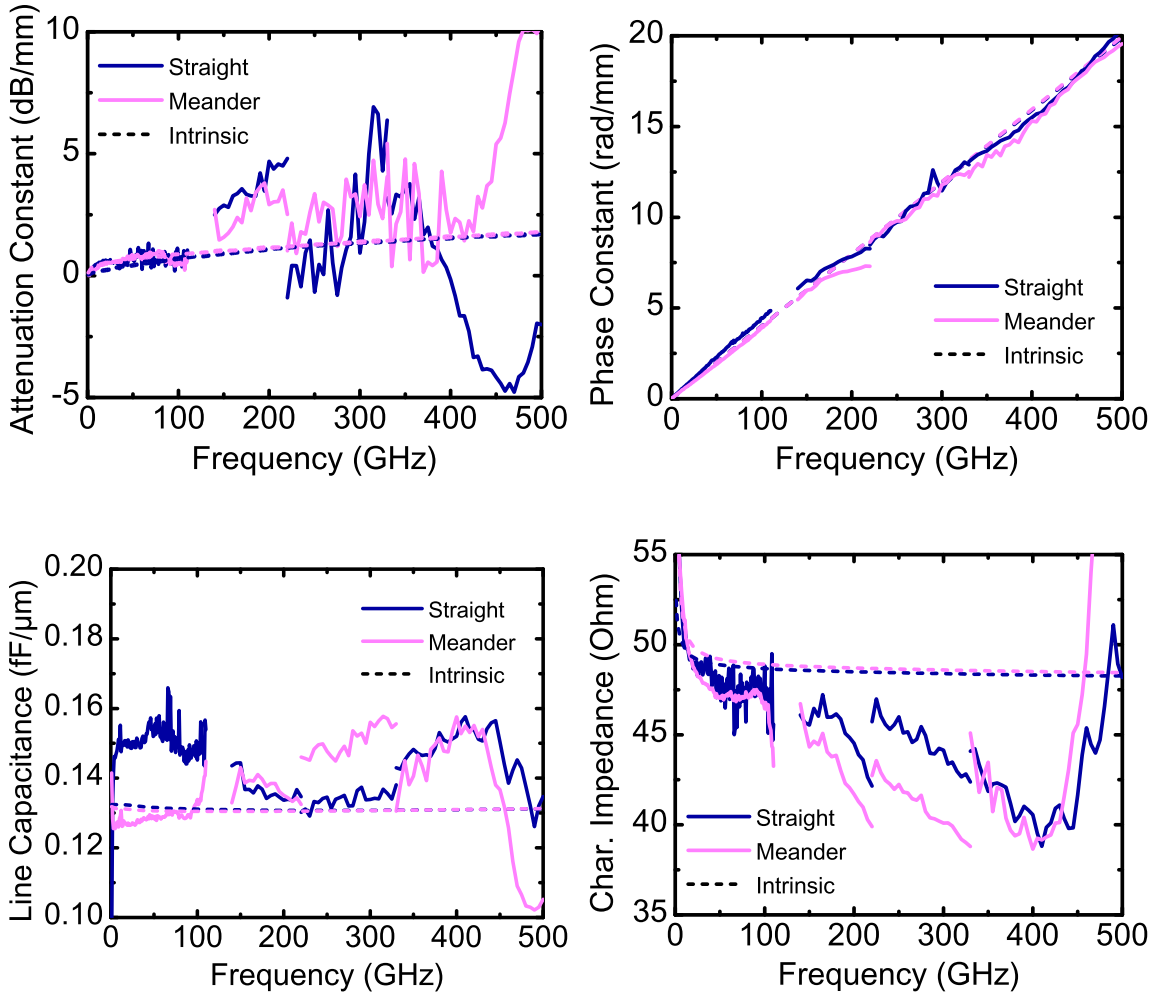


Figure 4.22: Measured electric parameters of L-500G and L-500G (M), marked as "straight" and "meander", respectively. Comparison with the intrinsic simulation of the same lines.

frequency bands. The port inductances of C-S (T-0) are symmetric and appear rather noisy, for this reason only L_2 is presented in Fig. 4.23.

Simulation well replicates the measured trends (slopes of the curves and slight difference between the two approaches, never above 3-4 pH). At high frequency, we expect by the simulated meander L_2 a separation of the meander approach from the intrinsic curve, which is not followed by the measurement in the 330-440 GHz range.

Moving on to the device analysis, in Fig. 4.24 the S-parameters of the HBT are depicted for the bias point where the f_T-peak is reached. Both simulated data (generated by the co-simulation approach) and measured data have been calibrated as before and de-embedded by our C-O/C-S structures.

As for the previous passive devices characterization, also in this meander- vs. straight-type design comparison only the thru and line standards for calibration are changing, whilst every other device data is the same.

We first consider the transmission coefficients. We can observe identical mag(S_{21})/arg(S_{21}), S_{21} being easy to characterize, but a slight deviation of both the meander curves starting from around 450 GHz can be also noticed.

mag(S_{12}) is well reproduced by the simulation curves particularly at low frequency. Trends are very similar for meander TRL and straight TRL curves, and confirmed by simulation. The curves slightly diverge around 440 GHz, though, yet the simulation does not fully capture this change. Similar considerations can be made for arg(S_{12}).

Surprisingly, the measurements after calibration with both meander or straight line show sim-

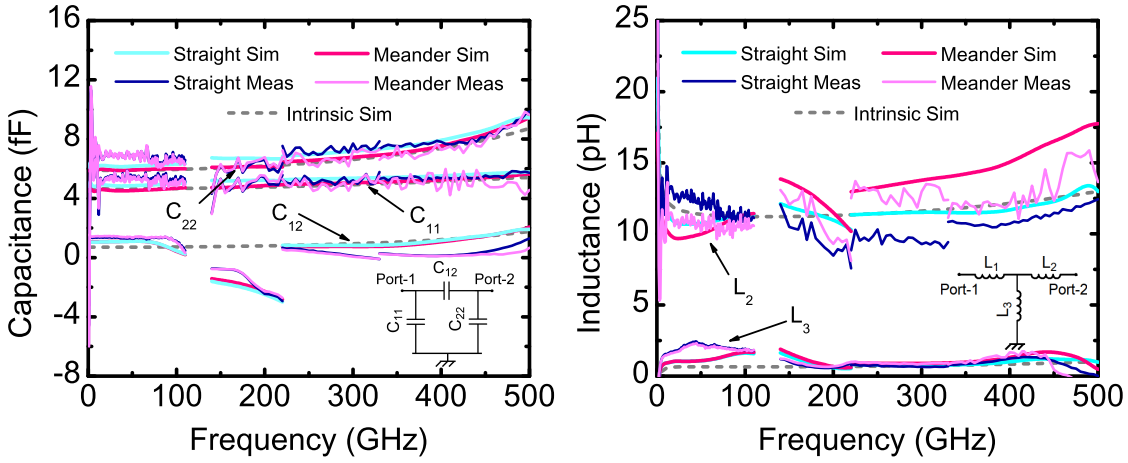


Figure 4.23: C-O (T-0) capacitances and C-S (T-0) inductances measurements and probe simulations (both classic and meander TRL calibration are applied) versus intrinsic simulation. As L_1 is symmetric to L_2, it is not shown.

ilar behaviors, that is different from the SPICE+EM co-simulation approach. Possible explanation can be physical effects inside the device that are not correctly captured by the compact model (e.g. NQS effects).

Let us now study the reflection coefficients. Simulation curves in mag(S_{11}) follow the measurement and the offset between straight and meander is reproduced. The curves' crossing which leads to a peak of the meander curve around 480 GHz is also replicated by simulation. However, the straight TRL curve also rises starting from 420 GHz, suggesting that the peak does not belong to the meander structure but is reinforced by it.

Moreover, we can observe a change in the monotonous decrease of arg(S_{11}) in both curves, again stronger for meander. Analogously to mag(S_{11}), mag(S_{22}) follows the measurement with an offset and tends to a peak at high frequency (meander case, above 500 GHz). In the 140-220 GHz range, measurement and simulation curves have opposite trends (for the classic case only). The change in the slope of both straight and meander is also visible by simulation between 100 and 110 GHz. Similarly to mag(S_{22}), arg(S_{22}) presents a simulation-measurement offset and a growing trend at high frequency (meander case).

The HF trend change we witness on the S-parameters is partially linked to those we observed on the electric parameters. The frequencies at which the trends start to variate are difficult to establish precisely from the S-parameters, but those found by alpha are quite consistent with the transistor's curves (Fig. 4.22).

By intrinsic simulation, we observed a strong dip in the magnitude of S_{11} and S_{22} of L-500G (both the straight and meander): see the orange and blue curves in Fig. 4.25. These parameters are supposed to be ideally well below 0 dB (and indeed are at most -30 dB): however, at 350 GHz on the meander line and 420 GHz on the straight line, they show a resonance and the values further decrease below -40 dB.

If we compute λ from the observed phenomenon's frequencies, $\overline{f}$, i.e.:

$$\lambda = \frac{c_0}{\sqrt{\varepsilon_{R,\text{eff}}} \cdot \overline{f}} \tag{4.5}$$

we find wavelengths of approximately 375 µm (straight) and 450 µm (meander). We observe that these wavelengths correspond to around twice the lengths of the lines: 187 µm, close to 185 µm of L-500G and 225 µm, close to the middle path length of L-500G (M) (see Table 4.3).

In conclusion, at those frequencies, the two L-500G lines behave as half-wave microstrip resonators and in fact, mag(S_{21}) = mag(S_{12}) of the lines display a slight increase at the corresponding frequencies if compared to the linearly decreasing trends at lower frequencies, as a sign of the maximum transmission of the signals.

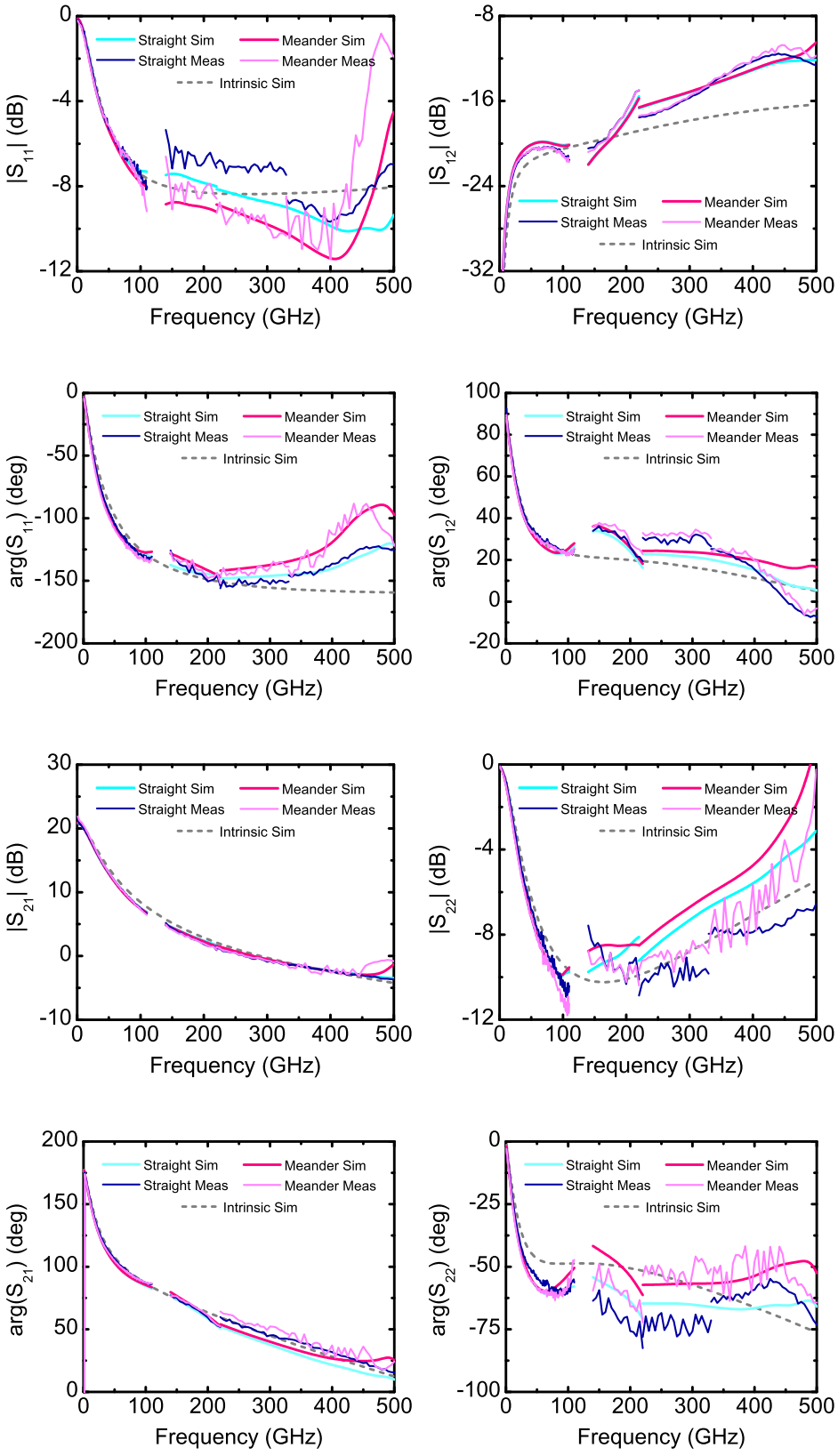


Figure 4.24: HBT's S-parameters measurements and transistor model+probe co-simulations (both classic and meander TRL calibration are applied) versus transistor model simulation (intrinsic) at $V_{CB} = 0\,V$, $V_{BE} = 0.9\,V$.

Due to the high frequency of occurrence, the S-parameters of the line are perturbed during measurement and when used to find the error terms of the TRL calibration, they may become a

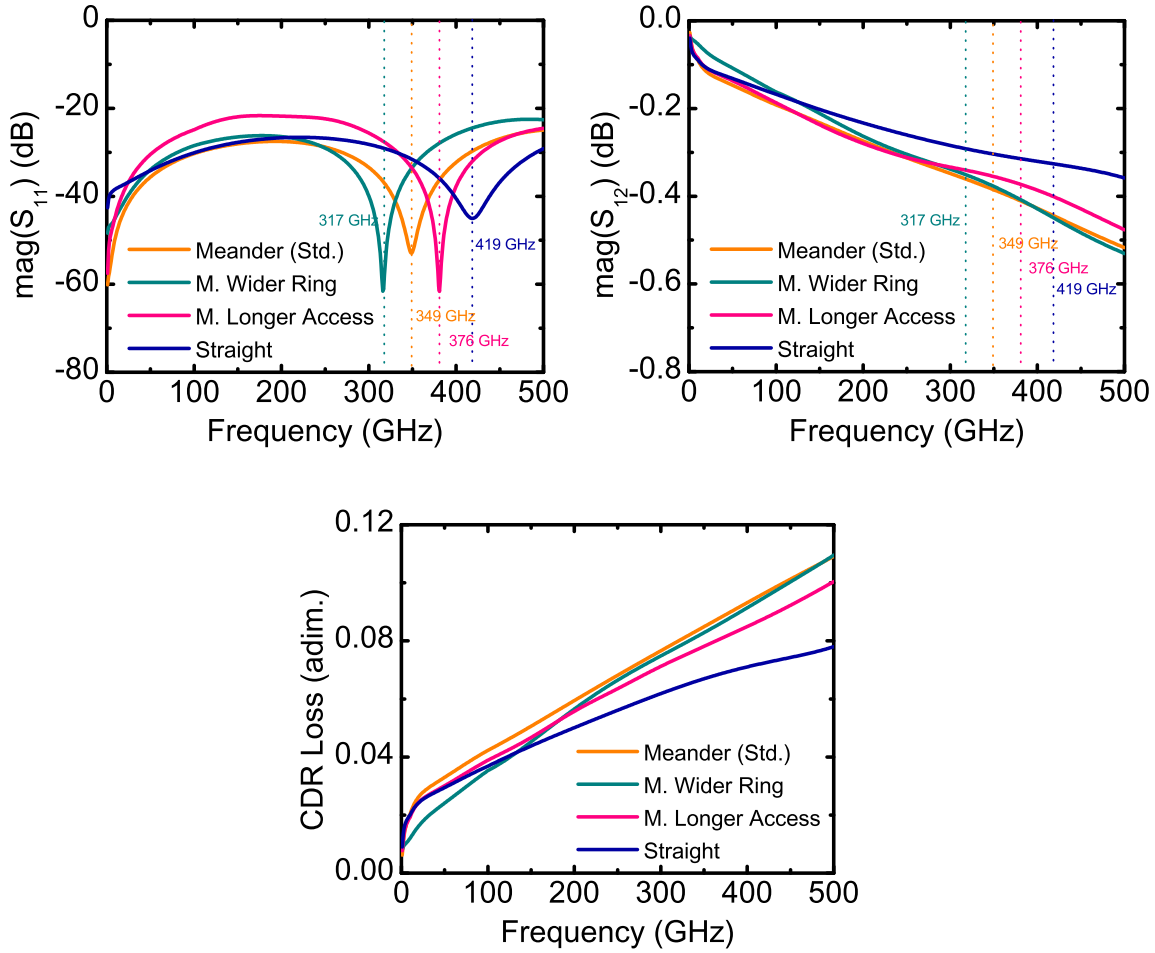


Figure 4.25: L-500G's S-parameters from intrinsic simulation (straight and meander): resonating frequencies are highlighted. Conduction-dielectric-radiation losses (CDR) are defined in [84] as CDR = $1 - |S_{11}|^2 - |S_{12}|^2$.

contributing factor to calibrated measurements' degradation.

As clearly depicted by Fig. 4.26, in fact, the electric field flowing over time through the intrinsic L-500G (M) at the resonating frequency stays on track following the meander pattern (the signal is generated at port 1, at the left of each figure).

At times t_6 and t_1, however, it can be seen bypassing the octagonal shape, and coupling with the next signal wave. The latency time during which the signal is covering the meander path is such that the next signal is generated when still the first has to reach port 2: due to the small separation between the access lines of the meander, the first signal "takes a shortcut" back to port 1.

To dive more into the coupling phenomenon and the resonating phenomenon, let us refer back to Fig. 4.25. The plots denoted by "M. Longer Access" and "M. Wider Ring" correspond to two modified versions of L-500G (M) that have been simulated to reduce crosstalk on the line and evaluate the impact of EM coupling (Fig. 4.27).

The first design is created by distancing the accesses to the quasi-octagonal pattern of L-500G from one another. The crosstalk area is therefore increased but the length of the meander line (both physical and effective) is reduced.

The "M. Wider Ring" design refers to an extension of the horizontal side of the quasi-octagonal pattern, to which we will refer as "ring". The port-to-port distance, hence the length of the line, is increased by 30 µm and the access are again set apart.

The electric field flux is shown in Fig. 4.28 for the "longer access" case and in Fig. 4.29 for the "wider ring" case. All the E-fields are shown at the same intensity scale (from 1 to 2 V/µm). Indeed,

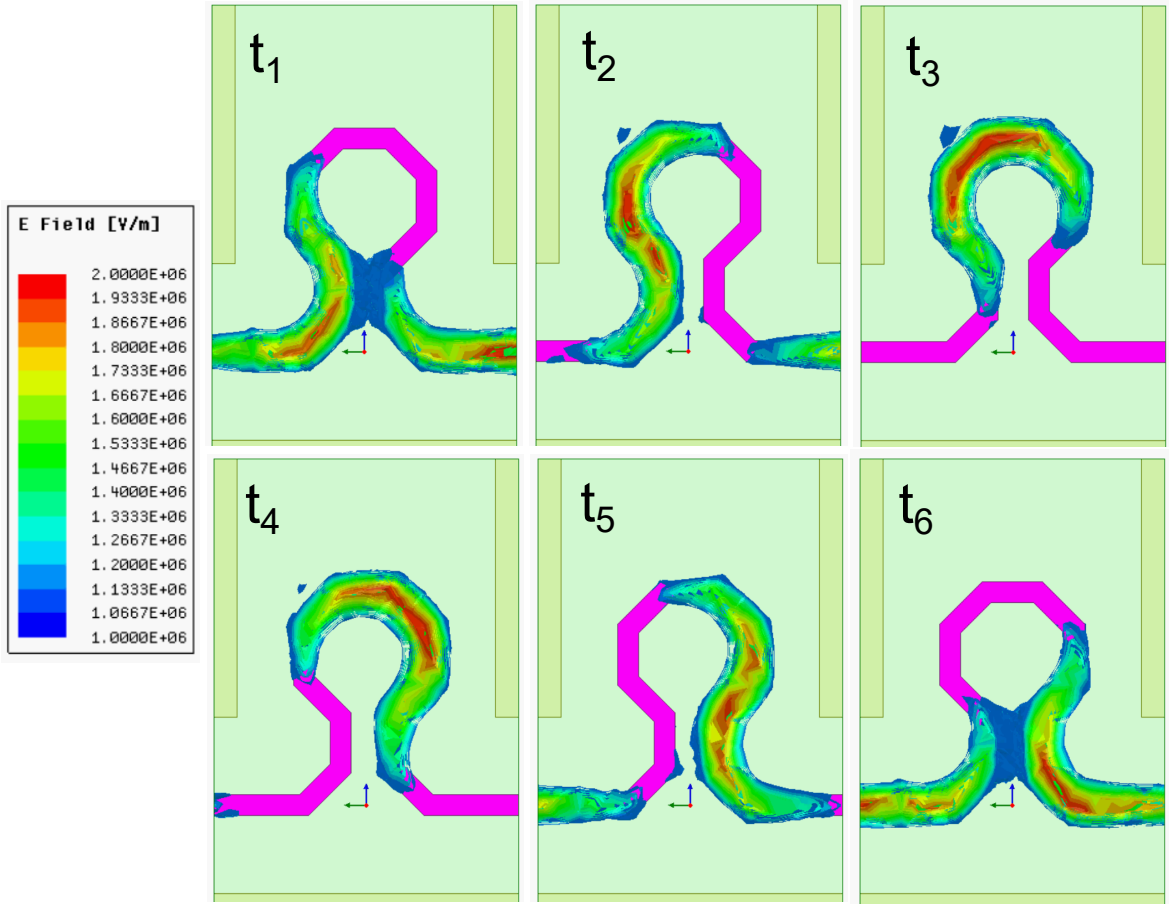


Figure 4.26: Electric field contour depicting the signal flow from port 1 to port 2 through L-500G (M) at 349 GHz at six different times. Bottom view.

the crosstalk is no more visible. However it is logically still present, only reduced in intensity.

If we look back to Fig. 4.25, we can finally observe that the dip at the resonating frequency on S_{11} moves proportionally to the length variation. With longer access, the total meander line is shorter and the dip moves to 376 GHz; inversely, with the wider ring, the total meander line is longer and the dip moves to 317 GHz.

On S_{12}, we can comment that the dip on the reflection coefficient indeed corresponds to higher energy transmission. While the rise on S_{12} around the resonating frequency is hardly visible in the case of the meander line and the "wider ring" meander, it is more clear on the straight line and the "longer access" meander line.

The current and Mason's gain, computed from the S-parameter we have previously seen, are plotted in Fig. 4.30. For f_T, measurements in the lower bands do not differ significantly, whilst simulation curves appear to indicate a few gigahertz difference.

Interestingly, f_{max} simulation and measurement do not match at lower frequency. This trend can be attributed to the measurement environment and not to the device itself. Reasonable results can be observed starting from 140 GHz, and in particular in the 220-360 GHz range.

The growth around 400 GHz (meander) and 425 GHz (straight) are predicted in simulation. Since the degraded trend of the calibrated measurement are also replicated by the probe simulation, hinting that this phenomenon is not (entirely, at least) linked to bad user's contact. We hypothesize that a line that reaches the $L = \lambda/2$ condition has a stronger bad impact on calibration than any possible crosstalk present on our novel meander line.

To complete this analysis of the transistor's figures of merit by adding the current gain $|H_{21}|$ and Mason's gain U, in Fig. 4.31 all of them are depicted for three different bias points.

The -20 dB/dec characteristic slope is visible for both of them, with very few differences be-

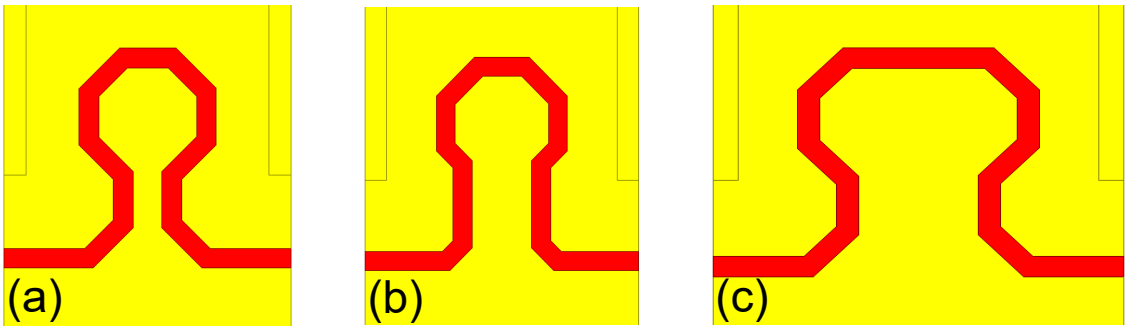


Figure 4.27: Variations on the meander L-500G (M): (a) standard meander, (b) longer access, (c) wider ring.

tween the straight- and meander-type approach. The current gain and the transit frequency curves show perfectly superimposed curves for every bias condition.

Most of these differences show up at the unilateral gain, and therefore at the maximum oscillation frequency, those being particularly difficult to evaluate. The differences are magnified by looking at f_{max}; nevertheless, the transistor's maximum frequency after both straight and meander calibration TRL approaches definitely look similar and the best frequency range for f_{max} determination starts from 220 GHz up to 325 GHz.

In conclusion, for the designed meander line, the validity range can be extended up to approximately 340 GHz, and the comparison should stop at the WR-3 band. Similarly, due to resonance in the straight line too, the classic TRL calibration with our toolkit should stop, as for that, at around 410 GHz. However, since the inverse peak on S_{11} associated to alpha (the Q factor) is lower than the meander case, we believe that the effect on the calibrated measurement is less important.

In order to extend this frequency range, we can think on inserting another –shorter– line for higher frequencies: with the meander design we gained in wafer area that can be now dedicated to a much less bulky third line (same area occupancy as the thru).

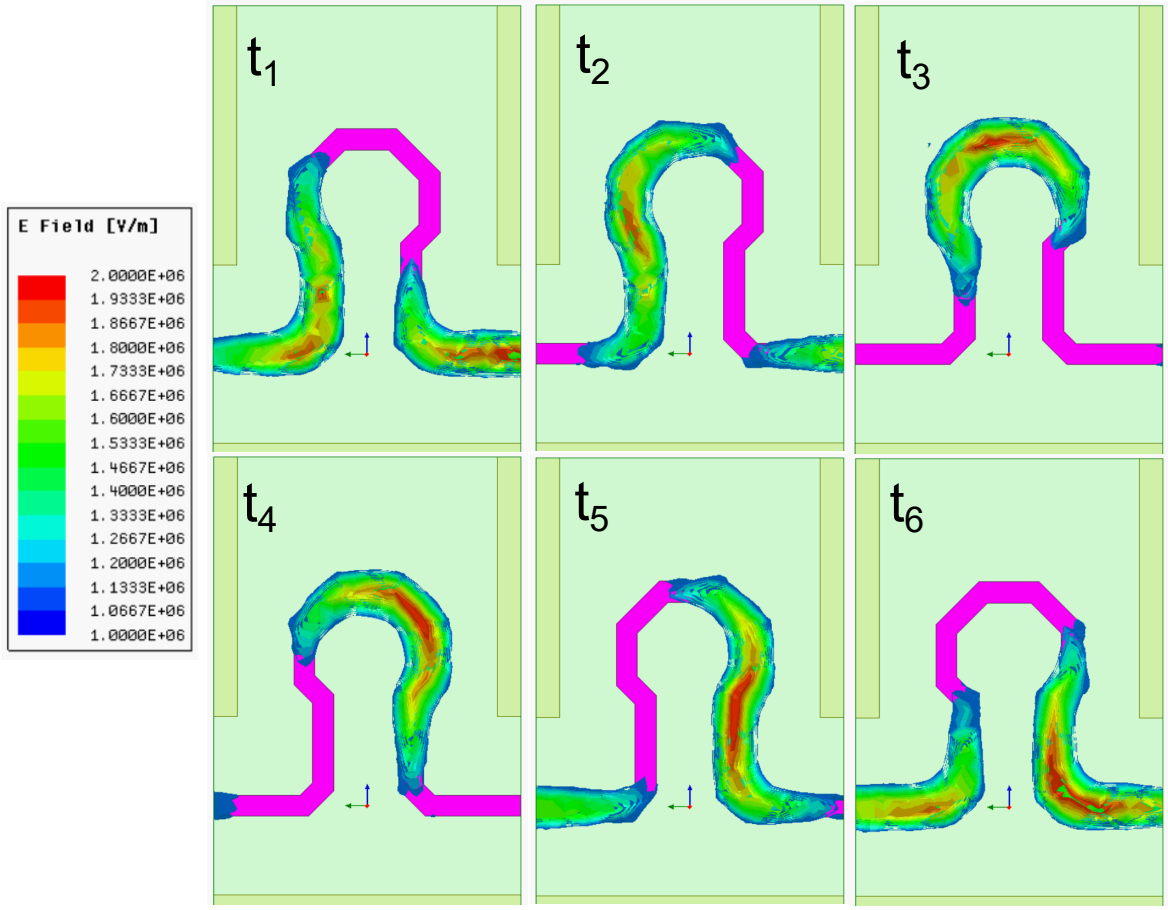


Figure 4.28: Electric field contour depicting the signal flow from port 1 to port 2 through a modified L-500G (M) at 376 GHz at six different times. Bottom view.

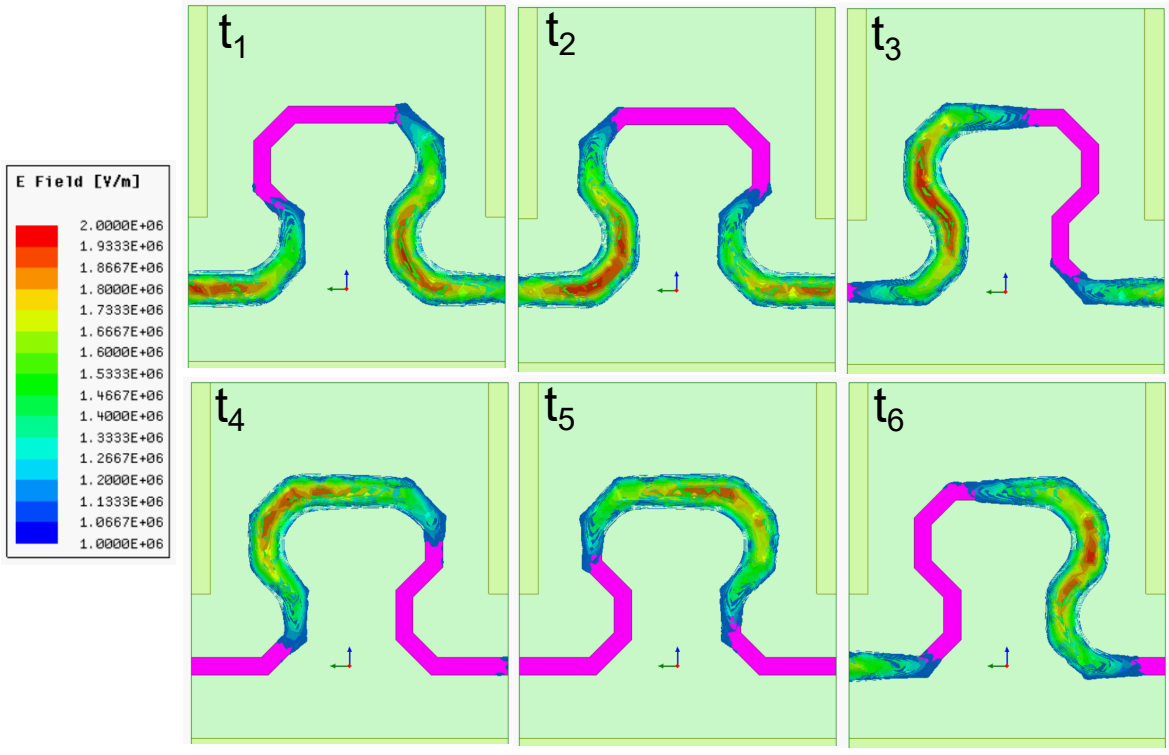
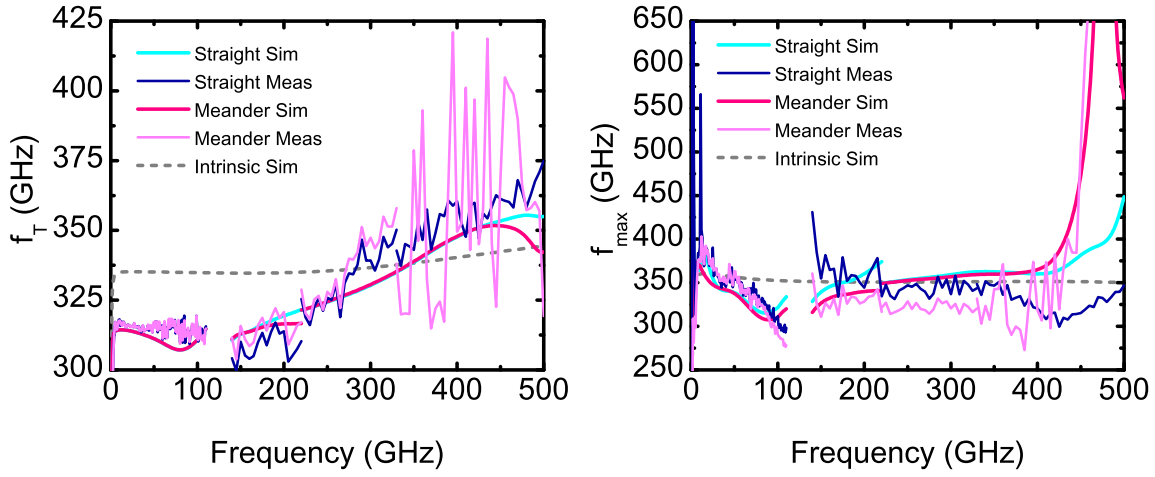


Figure 4.29: Electric field contour depicting the signal flow from port 1 to port 2 through a modified L-500G (M) at 317 GHz at six different times. Bottom view.



(a) Transit frequency f_T

(b) Max. oscillation frequency f_{max}

Figure 4.30: HBT's f_T and f_{max} measurements and transistor model+probe co-simulations (both classic and meander TRL calibration are applied) versus transistor model simulation (intrinsic) at $V_{CB} = 0\,V$, $V_{BE} = 0.9\,V$.

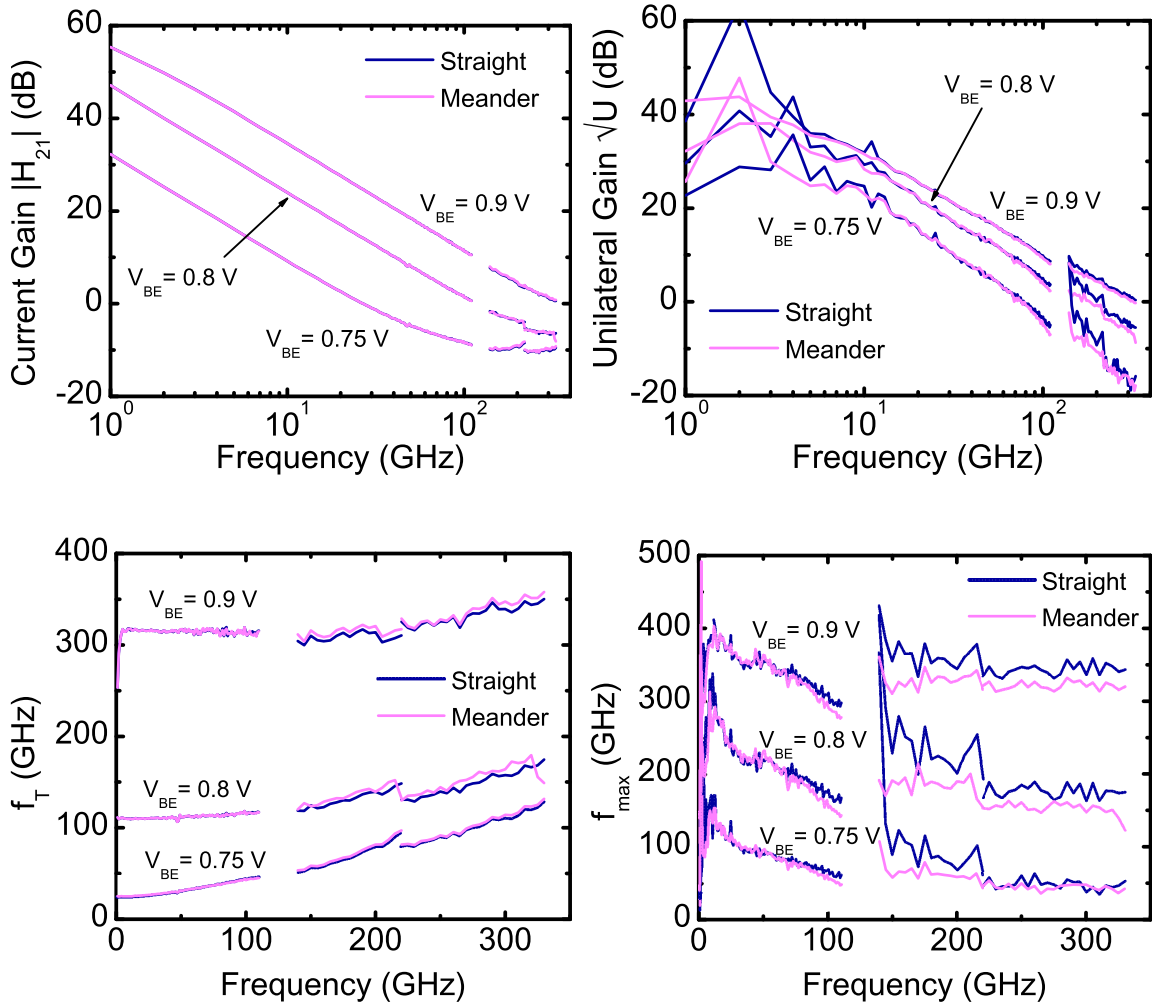


Figure 4.31: Main figures of merit of the HBT measured for different bias points ($V_{CB} = 0V$, $V_{BE} = 0.75, 0.8, 0.9\,V$).

4.3 Overview of Production Run 3

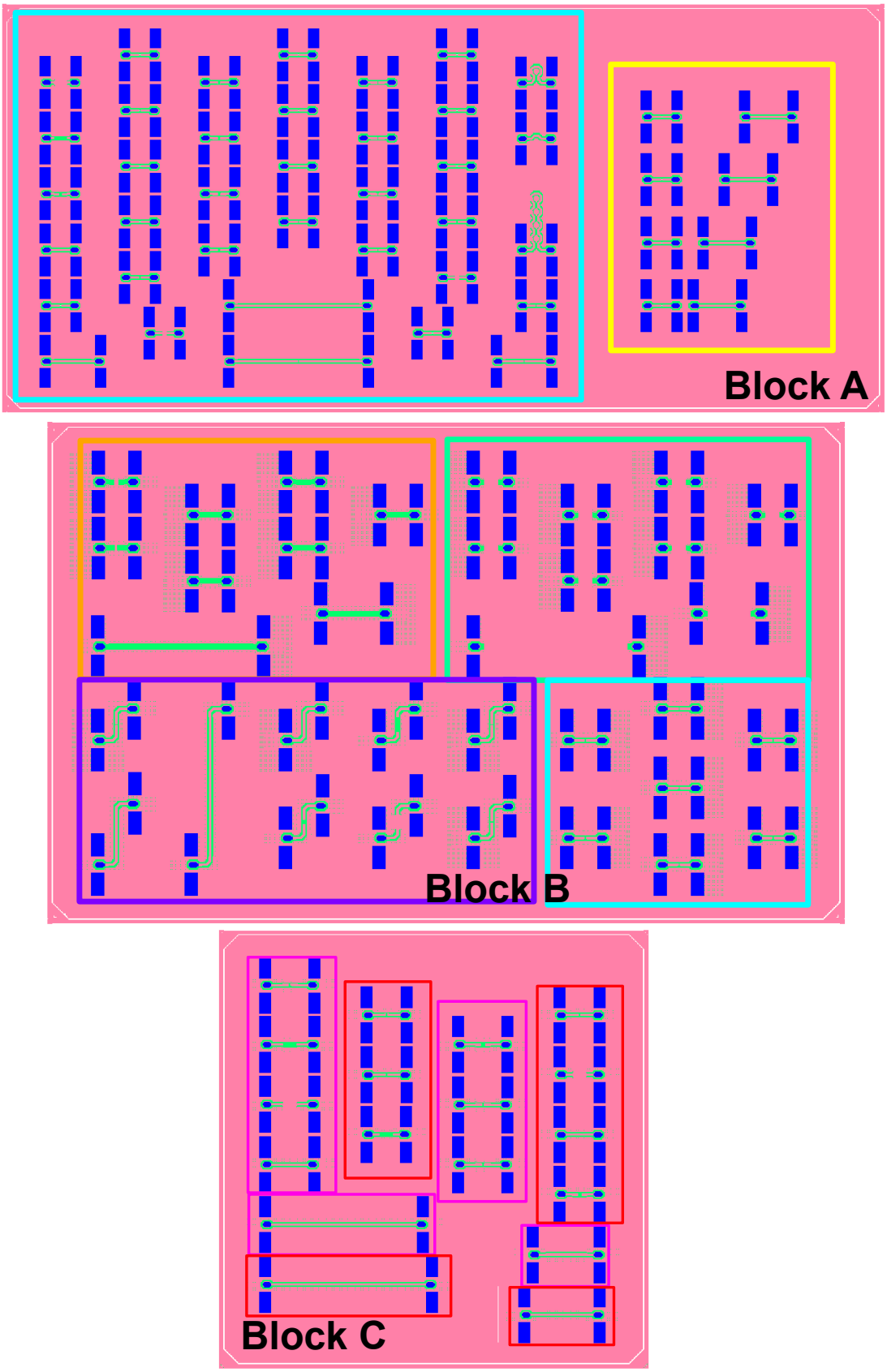


Figure 4.32: Floorplan artwork of production run 3 test structures: 3 blocks and different groups.

The continuous improvement led by the verification through measurement of calibration test structures brought us to recently design a newer run, tightly based on run 2 (the pad shield and the continuous ground plane, as well as all the previously characterized structures, make their return). The novel "run 3" (made of three separate blocks) is presented in Fig. 4.32.

Block A embeds a group of structures for "standard TRL" (framed in cyan): classic TRL calibration standards, de-embedding structures, HBTs, just like those of run 2, with the reference transistor (T-0) inserted multiple times in different die positions to evaluate differences due to process and neighbors' influence; meander TRL (lines and thru), with no new addition to what was presented in this work. Also in block A, surrounded by a yellow frame, dedicated structures for the neighbors' impact on coupling have been included. They have been separated from all other test structures by more than 300 µm and the same two structures are positioned with decreasing mutual distance from top to bottom. These structures are the open-M8 (on the left) and a specifically designed line, 150 µm long. The decreasing lengths between the (in-line) signal pads are 300, 200, 100 and 50 µm. By measuring these 4 configurations, we hope drawing educated conclusions on coupling below the probe shadow with adjacent devices.

Block B includes 4 different groups of structures, three of which are brand new designs for calibration. Framed in orange is the "M3 TRL", with no modifications compared to those previously presented. Surrounded by the cyan square is the "SOLT-M1" group: the central column are three HBTs with different emitter dimensions, the four devices split into two side columns are a short, an open, a line and a thru, all located at bottom metal layer (M1) for calibrating the transistors directly at M1, aiming to give increased accuracy particularly at LF. Framed in purple are the "shifted-pads" structures, and in green the "M6 TRL": both are introduced in the section below.

Finally, block C is composed of two groups of TRL calibration standards, both conceived for probe-to-probe crosstalk evaluation: the access lines are extended from 15 µm at both sides to 30 and 50 µm at both sides, increasing the inter-probe distance and the total length of the thru from 95 µm to 125 and 165 µm, respectively. The "+30u" structures are framed in red, the "+50u" in violet.

We do not dispose of measured data on these structures yet. However, we show in Appendix D some of the simulation results which have motivated the implementation on run 3 of the "shifted-pads" and "M6 TRL". These results have been retrieved by simulation only, and the prototyping was made on a run 1-based topology; however, the final implementation follows the rules established by run 2 layout.

4.3.1 Shifted-Pads

The shifted-pads lines implement the microstrip line in the in-plane perpendicular direction with respect to the standard line. In Fig. 4.33 we show the thru, as indicative of all other microstrip line shifted-pads implementations. From the accesses to the central portion of the line, the transition is made via a 45 degreed bend similar to the one of the unit cell, from the meander lines. The total line length (central path) is 210 µm, the central straight part being 98 µm long.

This design has previously been tested by HFSS simulation on a run 1-like layout, to benchmark its capability to follow similar trends than the standard run 1 measurement and to assess the deviation from the corresponding "standard TRL" simulation. Both measurement and simulation have been carried out with PP-110 and PP-220 probe sets up to 220 GHz. The results, shown in Fig. D.1, Appendix D, indicate that the use of a shifted-pads DUT with shifted-pads TRL calibration leads to results comparable to the standard.

We include also the EM field distributions in the complete probe setup of run 1 and run 1 modified with the shifted-pads (Fig. D.2, Appendix D, no neighbors but oxide ring, hence the field penetration below the DUT in figure). The field draws similar shape in both cases, and differences come from the surrounding environment (different probe coupling, linked to a field phase change).

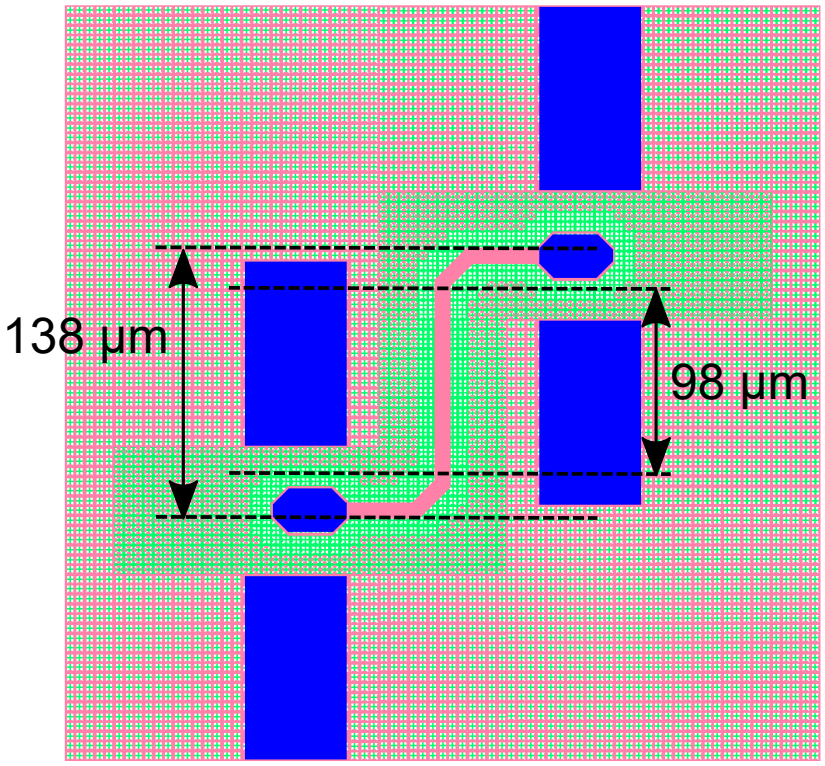


Figure 4.33: Top view of run 3 shifted-pads thru, with indicated lengths.

4.3.2 M6 TRL

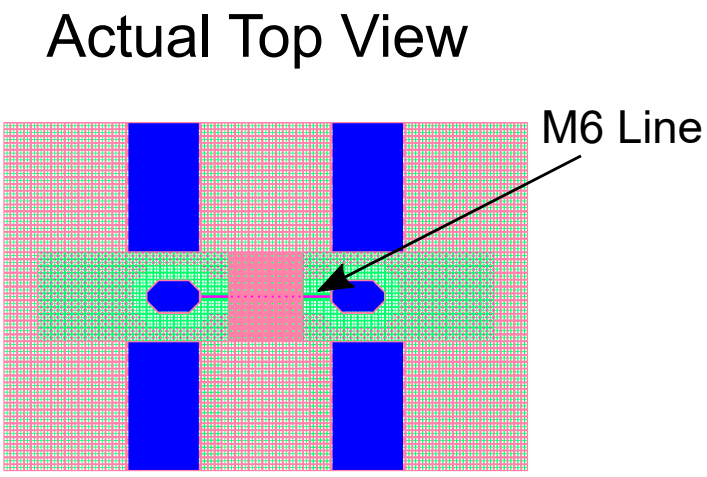


Figure 4.34: Top view of run 3 M6 thru, showing the microstrip line buried at M6.

The next innovative structures are based on the M6 line. We observe the M6 thru in Fig 4.34. Seen from above, the inter-probe distance is not changed, yet a metal block built on M8 connects the two portions of the side ground, covering the line, which is drawn on another level: the M6 level, located below. In the same figure we can see that this line comes as a thinner line, to preserve the 50-ohm Z_0 condition. Let us show in Fig 4.35 a 3D model implementation based on a pseudo-run 1 design, with the dimensions of the stripline indicated on it. The line is effectively "sandwitched" between the ground at M8 and the usual M1 ground.

Once again, these models are used to a complete probe simulation with DUT: C-O in Fig D.3, Appendix D, and the E-field in Fig D.4, Appendix D. Very good (flat) capacitances are drawn with the new M6 TRL, and field distribution is almost identical either with standard run 1 and M6 TRL.

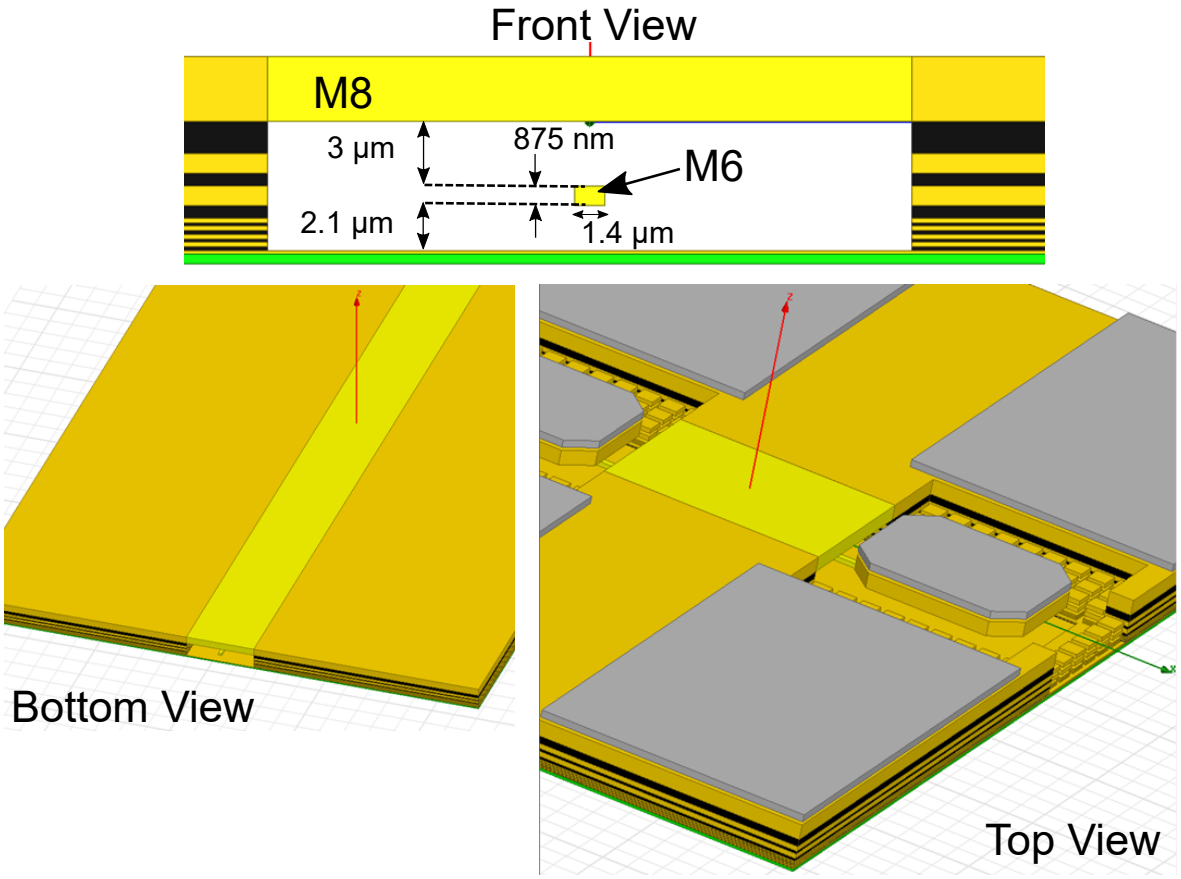


Figure 4.35: Detail of the asymmetric stripline implemented at M6 on its prototype version (with run 1-based topology).

Chapter 5

Conclusion

WE HAVE PRESENTED several techniques for sub-millimeter wave on-wafer calibration and device characterization.

Nowadays, the trends of sub-millimeter electronics are to close the "terahertz gap" in meaningful ways. We have seen that the investigative thrust in the field of terahertz electronics is leading to the design of compact systems that different fields can benefit from: the pharmaceutical sector, public safety, cancer research, autonomous cars, imaging, to name a few.

From the physical point of view, we have understood the interest of bipolar heterojunction transistors, which, thanks to bandgap engineering, allow high current gains and levels of amplification. We have defined two critical figures of merit of high frequency transitors, which thanks to the improvement of the Si-based HBT manufacturing techniques have constantly improved: the transit frequency f_T and the maximum oscillation frequency f_{max}. The first is related to the current gain and the transit time of the electrons between emitter and collector, the second describes the ability of a device integrated in a passive network to amplify a signal.

SiGe HBTs, which exhibit low band discontinuities and high integrative potential with CMOS, have been presented. STMicroelectronics' BiCMOS technology, B55, has been introduced. It has numerous advantages, including: low power consumption, high amplification, high current gain, increased packaging density. We have seen that ST's next BiCMOS transistor is already available: BiCMOS55X, which will be included in the next production run, ready to be characterized.

We learned that the power waves contain the same information as the incident and reflected voltages on a measurement structure, but are generated and received by the VNAs, which finally calculate the S parameters, which are the basis of the RF analysis of the devices. Thus, the architecture of a VNA has been studied and this allowed us to introduce the concept of measurement error, linked to non-ideal contributions that limit measurement accuracy. Errors are essentially of two types: random (due to noise, wrong contact, repetitiveness, etc ...) and systematic, the latter correctable with a calibration. For the S-parameter measurements, two calibration algorithms have been considered: SOLT (based on a 12-term error model), for which the knowledge of all the parameters of the calibration standards is a necessary condition, and TRL, which instead exploits the redundancy of calibration standards to minimize constraints on their characteristics. The latter is commonly used at high frequency, since it reduces systematic errors and allows to position the reference plane directly on the line. However, BEOL leaks remain a critical factor for the success of a TRL calibration. Several architectures of RF probes have been presented, as well as the importance of designing small sized RF pads, which help containing the onset of parasitic effects: in fact they can proportionally increase the measurement errors. The design choices are therefore critical, often more than the choice of the type of calibration algorithm, to obtain quality measurements. We also evaluated that the choice of a calibration on a substrate other than that of the DUT ("off-wafer") increases the measurement uncertainty as the validity of the calibration is reduced as the surrounding environment changes.

Our approach is therefore that of a completely on-wafer calibration, which extends up to 500 GHz with a reduced number of manipulations and data processing. For this reason, despite several

de-embedding techniques have been proposed, we have tried to verify the validity of a lumped approach, based only on an open and a short one, to push the reference plane to the transistor level, and thus remove all the parasitic contributions. Measurement setup considerations have been made. Signal transmission through suitable cables (coaxial or WG), extending the frequency of the VNAs and generating a convenient power level, to avoid instability and power drift.

Subsequently, the two production runs were extensively presented. The main differences of run 2, that is the continuous ground plane, the pad shield and the chessboard configuration have been presented as solutions to reduce the interference of the surroundings during the measurements and to improve the calibration process. The recommendations for the design were taken into account: increase of the inter-probe distance, of the inter-structure distance and reduction of the surface of the pads, among others. Furthermore, with a microstrip topology we limit the generation of higher order modes and at the same time, thanks to the reduced dimensions of our line, we reduce the parallel-plate modes and prevent, through the continuous ground plane, the propagation in the silicon substrate below. Our whole approach, in this manuscript, has been based on observing S-parameters and, where applicable, derived figures of merit, and we considered less clear and appealing, although already treated in the past of our research studies, to give insights on the error terms associated to the two runs. In addition to our measurement setup, we have also presented the different electromagnetic simulation methods: intrinsic, complete-probe and co-simulation, an innovative hybrid approach HICUM + probe tip model with the environment.

Finally we have presented our setup for data processing, our toolkit, the implementation of the SOLT and TRL calibration algorithms with impedance correction through the lumped-load method. We have seen to what extent the hypotheses for the use of this method are valid and proposed further simplifications to derive directly and automatically the error matrix linked to the characteristic impedance of the line whose value must be corrected. The simplification that provides the best results involves the use of a constant value of the line capacitance and the beta phase constant extracted at low frequency, but it also requires the attenuation constant, alpha, linked to the imaginary part of Z_0 and which is needed to correct the losses on our lines.

The comparison of the ISS SOLT, on-wafer SOLT and on-wafer TRL algorithms up to 500 GHz on three test structures allowed us to detect excellent results for the on-wafer SOLT (plus de-embedding), despite the often mentioned limitations of this calibration, showing difficulties in replicating the expected high frequency trend of the complete-short alone. SOLT calibration on active devices is comparable at low frequency (first band), slightly underestimates in the following ones. In the higher frequency band (330-500 GHz), the problems encountered with the TRL calibration do not allow us to draw firm conclusions. Also, we have seen that our on-wafer SOLT calibration is still non-ideally implemented since it needs an additional (pad-open + pad-short) de-embedding. Even though new preliminary data treatment showed that only pad-open is necessary, one might ask how to set the reference plane in the same position as TRL with no de-embedding in the first step of the calibration: this is still an open question.

The comparison between the two runs, on the other hand, allowed us to draw conclusions especially regarding the choices to be made in the design of test structures at millimeter-wave frequencies. Our continuous ground plane allows to isolate the probes from the silicon substrate and prevent substrate-to-substrate modes from propagating (causing oscillatory behavior, as seen in the coupling capacity of the pseudo-run 1 with an oxide ring). The absence of structures in line with the DUT, thanks to the chessboard configuration, limits the impact of the coupling on the neighbors, which would be difficult to remove given the limited control over the design of the surrounding environment, and which manifests itself, as we have seen, by an oscillatory behavior on the port capacitances. Finally, the metal ground plane makes it easier to remove probe-to-substrate parasitic modes through calibration, which are considered the main cause of measurement uncertainty. A common rule in industry of using a simple dielectric separating the devices (albeit by a chessboard configuration) does not protect, as seen through the EM simulation, from the numerous parasitic couplings.

Finally, in an effort to find different ways to perform an on-wafer TRL calibration, we have pre-

sented some innovative calibration structures. The first, the M3 layout, tries to bring the reference plane closer to the transistor, both vertically and horizontally. The analysis with a dedicated M3 calkit allowed us to explore and make broader considerations. The first notable conclusion is that the losses on the M3 line are not negligible, given the size of the strip. Secondly, the parasitic capacitances/inductances are better defined and constant in all frequency bands with a reference plane close to the DUT, without any accuracy loss for the calibration with respect to our convention: reference plane close to the pads and positioned at the end of the pad-open. The M3 layout demonstrates that the greatest parasitic contribution comes from the last stack levels (M1 specifically) but in general the parasites are well replicated. The M3 layout allows us to get rid of de-embedding and we were able to compare the different variants of TRL on active devices. We have highlighted a dependence of the measured curves by many factors, including the base (contact) resistance and the collector current, in addition to the aforementioned input power of the transistor.

The meander lines were then introduced: with their modular topology starting from unitary cells routed at 45 degrees, they allow to reduce the user's interventions by moving and re-arranging the probes during a measurement campaign. The actual lengths of these lines, needed to apply the algorithm, have been found through multiple methods, which showed negligible differences between them. The analysis of the transistor measurements has led us to highlight some resonances on the lines (straight and meander, although at a lower frequency for the meanders) which affect the final quality of the calibration.

Finally, run 3 has been presented, for an imminent characterization with innovative TRL calibration standards (M6 TRL and shifted pads) and embedded with structures that will allow to better understand the phenomena of crosstalk and interference with the surrounding environment.

Perspectives

In view of future improvements of our on-wafer characterization measurements, several perspectives are offered. One of the main concerns is the supplied power during HF measurements. We have seen numerous power problems during our measurement campaigns: the trace of the power injected into the device will be stabilized in frequency thanks to a bolometer. We will therefore be able to have greater control over the entire power band for the measurements not only of the active devices but also of the lines (to find exact values of the propagation constant, for example). We also plan to take full advantage of the probes we have acquired for high frequency characterization measurements. In particular, Dominion's TW-500 probe, given its architecture, appears to be a promising choice for excellent measurements in the 330-500 GHz band. Furthermore, we will use new 3D probe tip models we already have designed in our next run measurements: we have developed a model based on tomographic images of the IP-110 probe which allows very accurate simulations. The long-term goal is to extend these advanced models to all other probe sets we have.

In order to further improve measurement quality, as studied on the M3 layout, fabricating gold contacts (which allow for greater repeatability and less measurement deviation) might be a viable solution. In addition, a direct collector current control circuit, in place of the base voltage, should ensure less variation of the current in the transistor. In conclusion, these two solutions should lead to better measurements of the figures of merit, which are very sensitive to minimal parametric variations.

Alternatively, 3D TRL shows promising results as a one-tier calibration technique. A more detailed study of the signal dispersion in the final transition at the transistor level and the correction of the coupling capacitance would allow to better understand the potential of these lines.

As for meander line, in conclusion, we have shown that with the current design, these innovative structures are not yet usable beyond 330 GHz, but that the modular design still allows measurements of equal accuracy as the "standard" ones.

Bibliography

[1] On-Wafer Microwave De-Embedding Techniques. In *Microwave Systems and Applications*. URL http://www.intechopen.com/books/microwave-systems-and-applications/on-wafer-microwave-de-embedding-techniques. 30

[2] T. Agarwal. BiCMOS Technology: Fabrication and Applications, 2015. https://www.elprocus.com/bicmos-technology-fabrication-and-applications/. 12

[3] C. Andrei, D. Gloria, F. Danneville, P. Scheer, and G. Dambrine. Coupling on-wafer measurement errors and their impact on calibration and de-embedding up to 110 GHz for CMOS millimeter wave characterizations. In *2007 IEEE International Conference on Microelectronic Test Structures*, pages 253–256, March 2007. doi: 10.1109/ICMTS.2007.374494. 81, 84

[4] Aritsu. Vectorstar broadband VNA ME7838A/E/D, 2018. URL https://www.anritsu.com/en-us/test-measurement/products/me7838a. 31

[5] R. A. Bailey, M. A. Ussery, and P. J. Vail. A neutron hardness assurance screen based on high-frequency probe measurements. *IEEE Transactions on Nuclear Science*, 23(6):2020–2026, December 1976. doi: 10.1109/TNS.1976.4328617. 22

[6] M. F. Bauwens, N. Alijabbari, A. W. Lichtenberger, N. S. Barker, and R. M. Weikle. A 1.1 THz micromachined on-wafer probe. In *2014 IEEE MTT-S International Microwave Symposium (IMS2014)*, pages 1–4, June 2014. doi: 10.1109/MWSYM.2014.6848607. 27

[7] J. Bazzi. *Caractérisation des transistors bipolaires à Hétérojonction SiGe à très hautes fréquences*. PhD Thesis, Université Bordeaux I, 2011. 21

[8] J. Böck, K. Aufinger, S. Boguth, C. Dahl, H. Knapp, W. Liebl, D. Manger, T. F. Meister, A. Pribil, J. Wursthorn, R. Lachner, B. Heinemann, H. Rücker, A. Fox, R. Barth, G. Fischer, S. Marschmeyer, D. Schmidt, A. Trusch, and C. Wipf. SiGe HBT and BiCMOS process integration optimization within the DOTSEVEN project. In *2015 IEEE Bipolar/BiCMOS Circuits and Technology Meeting - BCTM*, pages 121–124, Oct 2015. doi: 10.1109/BCTM.2015.7340549. 13, 42

[9] C. Beng Sia. Minimizing discontinuities in wafer-level sub-THz measurements up to 750 GHz for device modelling applications. In *2017 89th ARFTG Microwave Measurement Conference (ARFTG)*, pages 1–4, June 2017. doi: 10.1109/ARFTG.2017.8000843. 105

[10] Gregory G. Boll and Harry J. Boll. Integrated circuit probing apparatus including a capacitor bypass structure, December 1994. URL https://www.freepatentsonline.com/5373231.html. 33

[11] J. Browne. *The Essentials of Vector Network Analyser*. Anritsu, 2009. 18

[12] R. Campbell, M. Andrews, L. Samoska, and A. Fung. Membrane tip probes for on-wafer measurements in the 220 to 325 ghz band. 01 2007. 33

[13] R. L. Campbell, M. Andrews, T. Lesher, and C. Wai. 220 GHz wafer probe membrane tips and waveguide-to-coax transitions. In *2005 European Microwave Conference*, volume 2, pages 4 pp.–1006, October 2005. doi: 10.1109/EUMC.2005.1610098. 33

[14] Q. Cassar. *Terahertz imaging and spectroscopy for breast cancer detection.* PhD Thesis, Université de Bordeaux, 2020. 4

[15] Q. Cassar, A. Alibadi, L. Mavarani, P. Hillger, J. Grzyb, G. MacGrogan, T. Zimmer, U. Pfeiffer, J.-P. Guillet, and P. Mounaix. Pilot study of freshly excised breast tissue response in the 300 – 600 ghz range. *Biomedical Optics Express*, 9:2930, 07 2018. doi: 10.1364/BOE.9.002930. 5

[16] P. Chevalier, G. Avenier, G. Ribes, A. Montagné, E. Canderle, D. Céli, N. Derrier, C. Deglise, C. Durand, T. Quémerais, M. Buczko, D. Gloria, O. Robin, S. Petitdidier, Y. Campidelli, F. Abbate, M. Gros-Jean, L. Berthier, J. D. Chapon, F. Leverd, C. Jenny, C. Richard, O. Gourhant, C. De-Buttet, R. Beneyton, P. Maury, S. Joblot, L. Favennec, M. Guillermet, P. Brun, K. Courouble, K. Haxaire, G. Imbert, E. Gourvest, J. Cossalter, O. Saxod, C. Tavernier, F. Foussadier, B. Ramadout, R. Bianchini, C. Julien, D. Ney, J. Rosa, S. Haendler, Y. Carminati, and B. Borot. A 55 nm triple gate oxide 9 metal layers SiGe BiCMOS technology featuring 320 GHz f_T/ 370 GHz f_{max} HBT and high-Q millimeter-wave passives. In *2014 IEEE International Electron Devices Meeting*, pages 3.9.1–3.9.3, December 2014. doi: 10.1109/IEDM.2014.7046978. 13, 14, 41

[17] P. Chevalier, G. Avenier, E. Canderle, A. Montagné, G. Ribes, and V. T. Vu. Nanoscale SiGe BiCMOS technologies: From 55 nm reality to 14 nm opportunities and challenges. In *2015 IEEE Bipolar/BiCMOS Circuits and Technology Meeting - BCTM*, pages 80–87, October 2015. doi: 10.1109/BCTM.2015.7340556. 13, 14

[18] P. Chevalier, M. Schroter, C. Bolognesi, V. D'Alessandro, M. Alexandrova, J. Böck, R. Fluckiger, S. Fregonese, B. Heinemann, C. Jungemann, R. Lövblom, C. Maneux, O. Ostinelli, A. Pawlak, N. Rinaldi, H. Rucker, G. Wedel, and T. Zimmer. Si/sige:c and inp/gaassb heterojunction bipolar transistors for thz applications. *Proceedings of the IEEE*, PP:1–16, 03 2017. doi: 10.1109/JPROC.2017.2669087. 8, 11, 13

[19] H. Cho and D. E. Burk. A three-step method for the de-embedding of high-frequency s-parameter measurements. *IEEE Transactions on Electron Devices*, 38(6):1371–1375, June 1991. doi: 10.1109/16.81628. 87

[20] M. Cho, G. Huang, C. Chiu, K. Chen, A. Peng, and Y. Teng. A cascade open-short-thru (COST) de-embedding method for microwave on-wafer characterization and automatic measurement. *IEICE Transactions*, 88-C:845–850, 01 2005. doi: 10.1093/ietele/e88-c.5.845. 87

[21] Gilles Dambrine. Chapter 2 - millimeter-wave characterization of silicon devices under small-signal regime: Instruments and measurement methodologies. In Giovanni Crupi and Dominique M.M.-P. Schreurs, editors, *Microwave De-embedding*, pages 47 – 96. Academic Press, Oxford, 2014. ISBN 978-0-12-401700-9. doi: https://doi.org/10.1016/B978-0-12-401700-9.00002-1. 26

[22] A. Davidson, K. Jones, and E. Strid. LRM and LRRM calibrations with automatic determination of load inductance. In *36th ARFTG Conference Digest*, volume 18, pages 57–63, November 1990. doi: 10.1109/ARFTG.1990.323996. 25

[23] M. Deng, S. Fregonese, D. Céli, P. Chevalier, M. De Matos, and T. Zimmer. Design of Silicon On-Wafer Sub-THz Calibration Kit. In *2017 Mediterranean Microwave Symposium (MMS)*, pages 1–4, November 2017. doi: 10.1109/MMS.2017.8497073. 42

[24] M. Deng, T. Quémerais, S. Bouvot, D. Gloria, P. Chevalier, S. Lépilliet, F. Danneville, and G. Dambrine. Small-signal characterization and modelling of 55nm SiGe BiCMOS HBT up to 325GHz. *Solid-State Electronics*, 129:150 – 156, 2017. ISSN 0038-1101. doi: https://doi.org/10.1016/j.sse.2016.11.012. URL http://www.sciencedirect.com/science/article/pii/S0038110116302805. 31

[25] N. Derrier, Andrej Rumiantsev, and Didier Céli. State-of-the-art and future perspectives in calibration and de-embedding techniques for characterization of advanced SiGe HBTs featuring sub-THz fT/fMAX. pages 1–8, 09 2012. doi: 10.1109/BCTM.2012.6352639. 30, 72, 87

[26] W. R. Eisenstadt and Y. Eo. S-parameter-based ic interconnect transmission line characterization. *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, 15(4): 483–490, August 1992. doi: 10.1109/33.159877. 65, 89

[27] G. F. Engen and C. A. Hoer. Thru-Reflect-Line: An Improved Technique for Calibrating the Dual Six-Port Automatic Network Analyzer. *IEEE Transactions on Microwave Theory and Techniques*, 27(12):987–993, December 1979. doi: 10.1109/TMTT.1979.1129778. 24, 25, 66

[28] H.-J. Eul and B. Schiek. Thru-Match-Reflect: One Result of a Rigorous Theory for De-Embedding and Network Analyzer Calibration. In *1988 18th European Microwave Conference*, pages 909–914, September 1988. doi: 10.1109/EUMA.1988.333924. 25

[29] H.-J. Eul and B. Schiek. A generalized theory and new calibration procedures for network analyzer self-calibration. *IEEE Transactions on Microwave Theory and Techniques*, 39(4), April 1991. doi: 10.1109/22.76439. 25

[30] A. Ferrero and U. Pisani. Two-port network analyzer calibration using an unknown 'thru'. *IEEE Microwave and Guided Wave Letters*, 2(12), December 1992. doi: 10.1109/75.173410. 25

[31] FormFactor. Probes waveguide selection, 2018. URL https://www.formfactor.com/download/probe-selection-guide/?wpdmdl=2561&refresh=5f6384dd6f81b1600357597. 33

[32] S. Fregonese, M. De Matos, M. Deng, M. Potereau, C. Ayela, K. Aufinger, and T. Zimmer. On-Wafer Characterization of Silicon Transistors Up To 500 GHz and Analysis of Measurement Discontinuities Between the Frequency Bands. *IEEE Transactions on Microwave Theory and Techniques*, 66(7):3332–3341, July 2018. doi: 10.1109/TMTT.2018.2832067. 31, 53, 54, 56, 72

[33] S. Fregonese, M. Deng, M. De Matos, C. Yadav, S. Joly, B. Plano, C. Raya, B. Ardouin, and T. Zimmer. Comparison of On-Wafer TRL Calibration to ISS SOLT Calibration With Open-Short De-Embedding up to 500 GHz. *IEEE Transactions on Terahertz Science and Technology*, 9(1):89–97, January 2019. doi: 10.1109/TTHZ.2018.2884612. 28, 31

[34] S. Fregonese, M. Cabbia, C. Yadav, M. Deng, S. Ranjan Panda, A. Chakravorty, and T. Zimmer. Analysis of high frequency measurement of transistors along with electromagnetic and SPICE co-simulation. *IEEE Transactions on Electron Devices*, 2020. To be published. 62

[35] A. Fung, L. Samoska, D. Pukala, D. Dawson, P. Kangaslahti, M. Varonen, T. Gaier, C. Lawrence, G. Boll, R. Lai, and X. Mei. On-wafer s-parameter measurements in the 325–508 GHz band. *IEEE Transactions on Terahertz Science and Technology - TTHZ*, 2:186–192, 03 2012. doi: 10.1109/TTHZ.2011.2182369. 33

[36] A. K. Fung, D. Dawson, L. Samoska, K. Lee, C. Oleson, and G. Boll. On-wafer vector network analyzer measurements in the 220-325 ghz frequency band. In *2006 IEEE MTT-S International Microwave Symposium Digest*, pages 1931–1934, June 2006. doi: 10.1109/MWSYM.2006.249811. 33

[37] T. Gaier, L. Samoska, C. Oleson, and G. Boll. On-wafer testing of circuits through 220 GHz. In *Ultrafast Electronics and Optoelectronics*, page UThC5. Optical Society of America, 1999. doi: 10.1364/UEO.1999.UThC5. URL http://www.osapublishing.org/abstract.cfm?URI=UEO-1999-UThC5. 33

[38] L. Galatro and M. Spirito. Analysis of residual errors due to calibration transfer in on-wafer measurements at mm-wave frequencies. In *2015 IEEE Bipolar/BiCMOS Circuits and Technology Meeting - BCTM*, pages 141–144, October 2015. doi: 10.1109/BCTM.2015.7340569. 28, 29

[39] L. Galatro and M. Spirito. Millimeter-wave on-wafer TRL calibration employing 3-D EM simulation-based characteristic impedance extraction. *IEEE Transactions on Microwave Theory and Techniques*, 65(4):1315–1323, January 2017. doi: 10.1109/TMTT.2016.2609413. 53, 56, 65

[40] L. Galatro, A. Pawlak, M. Schroter, and M. Spirito. Capacitively Loaded Inverted CPWs for Distributed TRL-Based De-Embedding at (Sub) mm-Waves. *IEEE Transactions on Microwave Theory and Techniques*, 65(12):4914–4924, December 2017. doi: 10.1109/TMTT.2017.2727498. 31, 87

[41] Luca Galatro. *Advanced calibration and measurement techniques for (sub)millimeter wave devices characterization*. PhD thesis, TU Delft, 2019. 26, 27

[42] A. Gauthier, J. Borrelf, P. Chevalier, G. Avenier, A. Montagne, M. Juhel, R. Duru, L. Clement, C. Borowiak, M. Buczko, and C. Gaquière. 450 ghz f_t sige:c hbt featuring an implanted collector in a 55-nm cmos node. pages 72–75, 10 2018. doi: 10.1109/BCICTS.2018.8551057. 14

[43] E. M. Godshalk. Surface wave phenomenon in wafer probing environments. In *40th ARFTG Conference Digest*, volume 22, pages 10–19, December 1992. doi: 10.1109/ARFTG.1992.326994. 32

[44] G. Gronau. *Höchstfrequenztechnik*. Springer, 2001. ISBN 978-3-642-56620-2. 25

[45] M. S. Gupta. Power gain in feedback amplifiers, a classic revisited. *IEEE Transactions on Microwave Theory and Techniques*, 40(5):864–879, May 1992. doi: 10.1109/22.137392. 10

[46] B. P. Hand. Developing Accuracy Specifications for Automatic Network Analyzer Systems. *Hewlett Packard J*, 21:16–19, 1970. URL https://www.hpl.hp.com/hpjournal/pdfs/IssuePDFs/1970-02.pdf. 23

[47] M. Hiebel. *Fundaments of Vector Network Analysis*. Rohde & Schwarz GmbH & Co., 2008. ISBN 978-3-939837-06-0. 19

[48] GGB Industries Inc. Model 110H high performance microwave probes, 2020. URL https://ggb.com/wp-content/uploads/2017/06/mod110h.pdf. 33

[49] GGB Industries Inc. Model 220H high performance microwave probes, 2020. URL https://ggb.com/wp-content/uploads/2017/06/mod220.pdf. 33

[50] GGB Industries Inc. Model 325B high performance microwave probes, 2020. URL https://ggb.com/wp-content/uploads/2017/06/Model325B.pdf. 33

[51] GGB Industries Inc. Model 500B high performance microwave probes, 2020. URL https://ggb.com/wp-content/uploads/2017/06/Model500B.pdf. 33

[52] K.L. Kaiser. *Electromagnetic Compatibility Handbook*. Electrical engineering handbook series. Taylor & Francis, 2004. ISBN 978-0-8493-2087-3. URL https://books.google.td/books?id=nZzOAsroBIEC. 32

[53] I. M. Kang, S. Jung, T. Choi, J. Jung, C. Chung, H. Kim, H. Oh, H. W. Lee, G. Jo, Y. Kim, H. Kim, and K. Choi. Five-step (pad–pad short–pad open–short–open) de-embedding method and its verification. *IEEE Electron Device Letters*, 30(4):398–400, February 2009. doi: 10.1109/LED.2009.2013881. 87

[54] T. E. Kolding. A four-step method for de-embedding gigahertz on-wafer cmos measurements. *IEEE Transactions on Electron Devices*, 47(4):734–740, April 2000. doi: 10.1109/16.830987. 87

[55] M. C. A. M. Koolen, J. A. M. Geelen, and M. P. J. G. Versleijen. An improved de-embedding technique for on-wafer high-frequency characterization. In *Proceedings of the 1991 Bipolar Circuits and Technology Meeting*, pages 188–191, September 1991. doi: 10.1109/BIPOL.1991.160985. 30, 87

[56] H. Kroemer. Theory of a wide-gap emitter for transistors. *Proceedings of the IRE*, 45(11):1535–1537, November 1957. ISSN 0096-8390. doi: 10.1109/JRPROC.1957.278348. 7

[57] K. Kuhn, M. Agostinelli, S. Ahmed, S. Chambers, S. Cea, S. Christensen, P. Fischer, J. Gong, C. Kardas, T. Letson, L. Henning, A. Murthy, H. Muthali, B. Obradovic, P. Packan, S. W. Pae, I. Post, S. Putna, K. Raol, A. Roskowski, R. Soman, T. Thomas, P. Vandervoorn, M. Weiss, and I. Young. A 90 nm communication technology featuring SiGe HBT transistors, RF CMOS, precision R-L-C RF elements and 1 /spl mu/m2 6-T SRAM cell. In *Digest. International Electron Devices Meeting*, pages 73–76, December 2002. doi: 10.1109/IEDM.2002.1175782. 14

[58] Der Sun Lee and J. G. Fossum. Energy-band distortion in highly doped silicon. *IEEE Transactions on Electron Devices*, 30(6):626–634, June 1983. ISSN 0018-9383. doi: 10.1109/T-ED.1983.21181. 6

[59] Q. Liang, J. Cressler, G. Niu, Y. Lu, G. Freeman, David C. Ahlgren, R. Malladi, K. Newton, and D. Harame. A simple four-port parasitic deembedding methodology for high-frequency scattering parameter and noise characterization of sige hbts. *IEEE Transactions on Microwave Theory and Techniques*, 51:2165–2174, November 2003. doi: 10.1109/TMTT.2003.818580. 87

[60] S. M. J. Liu and G. G. Boll. A new probe for w-band on-wafer measurements. In *1993 IEEE MTT-S International Microwave Symposium Digest*, pages 1335–1338 vol.3, June 1993. doi: 10.1109/MWSYM.1993.277123. 37

[61] A. M. Mangan, S. P. Voinigescu, Ming-Ta Yang, and M. Tazlauanu. De-embedding transmission line measurements for accurate modeling of ic designs. *IEEE Transactions on Electron Devices*, 53(2):235–241, January 2006. doi: 10.1109/TED.2005.861726. 87

[62] P. Manuel, C. Raya, M. De Matos, S. Fregonese, A. Curutchet, M. Zhang, B. Ardouin, and T. Zimmer. Limitations of On-Wafer Calibration and De-Embedding Methods in the Sub-THz Range. *Journal of Computer and Communications*, 01:25–29, 2013. doi: 10.4236/jcc.2013.16005. 103

[63] R. B. Marks and D. F. Williams. Interconnection transmission line parameter characterization. In *40th ARFTG Conference Digest*, volume 22, pages 88–95, December 1992. doi: 10.1109/ARFTG.1992.327004. 65

[64] R.B. Marks. A multiline method of network analyzer calibration. *IEEE Transactions on Microwave Theory and Techniques*, 39(7), July 1991. doi: 10.1109/22.85388. 26

[65] R.B. Marks and D.F. Williams. Characteristic impedance determination using propagation constant measurement. *IEEE Microwave and Guided Wave Letters*, 1(6):141–143, June 1991. doi: 10.1109/75.91092. 65, 66

[66] Roger B Marks and Dylan F Williams. A General Waveguide Circuit Theory. *Journal of research of the National Institute of Standards and Technology*, 97(5):533–562, 1992. ISSN 1044-677X. doi: 10.6028/jres.097.024. URL https://pubmed.ncbi.nlm.nih.gov/28053445. Publisher: [Gaithersburg, MD] : U.S. Dept. of Commerce, National Institute of Standards and Technology. 26

[67] J. Marzouk, S. Arscott, A. El Fellahi, K. Haddadi, T. Lasri, C. Boyaval, and G. Dambrine. MEMS probes for on-wafer RF microwave characterization of future microelectronics: design, fabrication and characterization. *Journal of Micromechanics and Microengineering*, 25 (7):075024, jun 2015. doi: 10.1088/0960-1317/25/7/075024. 27

[68] J. W. Matthews. Defects associated with the accommodation of misfit between crystals. *Journal of Vacuum Science and Technology 12, 126*, 1975. 11

[69] X. Mei, W. Yoshida, M. Lange, J. Lee, J. Zhou, P. Liu, K. Leong, A. Zamora, J. Padilla, S. Sarkozy, R. Lai, and W. Deal. First demonstration of amplification at 1 thz using 25-nm inp high electron mobility transistor process. *Electron Device Letters, IEEE*, 36:327–329, 04 2015. doi: 10.1109/LED.2015.2407193. 8

[70] D. Müller, J. Schafer, D. Geenen, H. Massler, A. Tessmann, A. Leuther, T. Zwick, and I. Kallfass. Electromagnetic field simulation of mmics including rf probe tips. pages 900–903, 10 2017. doi: 10.23919/EuMC.2017.8230990. 61

[71] Daniel Müller. *RF Probe-Induced On-Wafer Measurement Errors in the Millimeter-Wave Frequency Range*. PhD Thesis, Karlsruher Institut für Technologie (KIT), 2018. 26

[72] R. S. Muller and T. I. Kamins. *Device Electronics for Integrated Circuits*. John Wiley & Sons Inc, 2003. 6, 10, 11

[73] NIST. Radio frequency and analog/mixed-signal technologies, January 2012. URL https://www.nist.gov/publications/radio-frequency-and-analogmixed-signal-technologies. 27

[74] D. J. Paul. Si/SiGe Heterojunction Bipolar Transistors, 2004. http://userweb.eng.gla.ac.uk/douglas.paul/SiGe/HBT.html. 11

[75] Ashish Y. Pawar, Deepak D. Sonawane, Kiran B. Erande, and Deelip V. Derle. Terahertz technology and its applications. *Drug Invention Today*, 5(2):157 – 163, 2013. ISSN 0975-7619. doi: https://doi.org/10.1016/j.dit.2013.03.009. URL http://www.sciencedirect.com/science/article/pii/S0975761913000264. 3

[76] U. R. Pfeiffer, E. Ojefors, A. Lisauskas, and H. G. Roskos. Opportunities for silicon at mmWave and Terahertz frequencies. In *2008 IEEE Bipolar/BiCMOS Circuits and Technology Meeting*, pages 149–156, October 2008. doi: 10.1109/BIPOL.2008.4662734. 4

[77] G. N. Phung, F. J. Schmückle, and W. Heinrich. Parasitic effects and measurement uncertainties in multi-layer thin-film structures. In *2013 European Microwave Conference*, pages 318–321, October 2013. doi: 10.23919/EuMC.2013.6686655. 55, 56, 81, 84

[78] G. N. Phung, F. J. Schmückle, R. Doerner, T. Fritzsch, and W. Heinrich. Impact of parasitic coupling on multiline trl calibration. In *2017 47th European Microwave Conference (EuMC)*, pages 835–838, October 2017. doi: 10.23919/EuMC.2017.8230974. 82

[79] G. N. Phung, F. J. Schmückle, R. Doerner, B. Kähne, T. Fritzsch, U. Arz, and W. Heinrich. Influence of microwave probes on calibrated on-wafer measurements. *IEEE Transactions on Microwave Theory and Techniques*, 67(5):1892–1900, March 2019. doi: 10.1109/TMTT.2019.2903400. 53, 54, 56, 81, 104

[80] M. Potereau, C. Raya, M. DeMatos, S. Fregonese, A. Curutchet, M. Zhang, B. Ardouin, and T. Zimmer. Limitations of on-wafer calibration and de-embedding methods in the sub-THz range. *Journal of Computer and Communications*, 01:25–29, November 2013. doi: 10.4236/jcc.2013.16005. 81

[81] M. Potereau, S. Fregonese, A. Curutchet, P. Baureis, and T. Zimmer. New 3d-trl structures for on-wafer calibration for high frequency s-parameter measurement. pages 167–170, 09 2015. doi: 10.1109/EuMC.2015.7345726. 101, 102

[82] M. Potéreau, M. Deng, C. Raya, B. Ardouin, K. Aufinger, C. Ayela, M. De Matos, A. Curutchet, S. Frégonèse, and T. Zimmer. Meander type transmission line design for on-wafer TRL calibration. In *2016 46th European Microwave Conference (EuMC)*, pages 381–384, October 2016. doi: 10.1109/EuMC.2016.7824358. 104, 105

[83] F. Pourchon, C. Raya, N. Derrier, P. Chevalier, D. Gloria, S. Pruvost, and D. Celi. From measurement to intrinsic device characteristics: Test structures and parasitic determination. In *2008 IEEE Bipolar/BiCMOS Circuits and Technology Meeting*, pages 232–239, October 2008. doi: 10.1109/BIPOL.2008.4662751. 87

[84] D.M. Pozar. *Microwave Engineering, 4th Edition*. Wiley, 2011. ISBN 9781118213636. URL https://books.google.fr/books?id=JegbAAAAQBAJ. 55, 71, 90, 112, VII

[85] C. Raya. *Modelisation et Optimisation de Transistors Bipolaires Heterojonction Si/SiGeC Ultra Rapides Pour Applications Millimetriques*. PhD thesis, Université de Bordeaux, 2008. 21

[86] G. M. Rebeiz. Millimeter-wave and terahertz integrated circuit antennas. *Proceedings of the IEEE*, 80(11):1748–1770, November 1992. ISSN 0018-9219. doi: 10.1109/5.175253. 3

[87] T. J. Reck, L. Chen, C. Zhang, C. Groppi, H. Xu, A. Arsenovic, N. S. Barker, A. Lichtenberger, and R. M. Weikle. Micromachined on-wafer probes. In *2010 IEEE MTT-S International Microwave Symposium*, pages 65–68, May 2010. doi: 10.1109/MWSYM.2010.5517580. 27, 36

[88] T. J. Reck, L. Chen, C. Zhang, A. Arsenovic, C. Groppi, A. Lichtenberger, R. M. Weikle, and N. S. Barker. Micromachined probes for submillimeter-wave on-wafer measurements—part ii: Rf design and characterization. *IEEE Transactions on Terahertz Science and Technology*, 1 (2):357–363, 2011. doi: 10.1109/TTHZ.2011.2165020. 27, 33

[89] T. J. Reck, L. Chen, C. Zhang, A. Arsenovic, C. Groppi, A. W. Lichtenberger, R. M. Weikle, and N. S. Barker. Micromachined probes for submillimeter-wave on-wafer measurements—part i: Mechanical design and characterization. *IEEE Transactions on Terahertz Science and Technology*, 1(2):349–356, 2011. doi: 10.1109/TTHZ.2011.2165013. 27, 33

[90] N. Rinaldi and M. Schröter. *Silicon-Germanium Heterojunction Bipolar Transistors for mm-Wave Systems: Technology, Modeling and Circuit Applications*. River Publishers Series in Electronic Materials and Devices. River Publishers, 2018. ISBN 9788793519619. URL https://books.google.fr/books?id=0O1qDwAAQBAJ. 7, 10, 12

[91] M. J. W. Rodwell, R. Yu, P. Reddy, S. Allen, and U. Bhattacharya. Active probes for on-wafer millimeter-wave network analysis. In *Proceedings of Conference on Precision Electromagnetic Measurements Digest*, pages 8–, June 1994. doi: 10.1109/CPEM.1994.333314. 27

[92] A. Rumiantsev and R. Doerner. Rf probe technology: History and selected topics. *Microwave Magazine, IEEE*, 14:46–58, 11 2013. doi: 10.1109/MMM.2013.2280241. 27, 32, 37

[93] A. Rumiantsev, P. Sakalas, F. Pourchon, P. Chevalier, N. Derrier, and M. Schroter. Application of on-wafer calibration techniques for advanced high-speed bicmos technology. In *2010 IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, pages 98–101, October 2010. doi: 10.1109/BIPOL.2010.5667929. 66, 87

[94] Andrej Rumiantsev, Paulius Sakalas, N. Derrier, Didier Céli, and M. Schroter. Influence of probe tip calibration on measurement accuracy of small-signal parameters of advanced bicmos hbts. pages 203–206, 10 2011. doi: 10.1109/BCTM.2011.6082782. 26

[95] Doug Rytting. An analysis of vector measurement accuracy enhancement techniques, 1980. URL http://na.support.keysight.com/faq/accuracy.pdf. 25

[96] Doug Rytting. Network Analyzer Error Models and Calibration Methods, 1996. URL http://emlab.uiuc.edu/ece451/appnotes/Rytting_NAModels.pdf. 21, 25

[97] Roberto S. Murphy. Characterisation of Semiconductor Devices in the High Frequency Regime, December 2019. URL http://mos-ak.org/silicon_valley_2019/presentations/10_Murphy_MOS-AK_Dec2019.pdf. 30

[98] A. M. E. Safwat, M. Andrews, L. Hayden, K. R. Gleason, and E. Strid. A probe technology for 110+ GHz integrated circuits with aluminum pads. In *59th ARFTG Conference Digest, Spring 2002.*, pages 60–66, June 2002. doi: 10.1109/ARFTGS.2002.1214682. 37

[99] A.M.E. Safwat and L. Hayden. Sensitivity analysis of calibration standards for fixed probe spacing on-wafer calibration techniques [vector network analyzers]. In *2002 IEEE MTT-S International Microwave Symposium Digest (Cat. No.02CH37278)*, volume 3, June 2002. doi: 10.1109/MWSYM.2002.1012323. 29

[100] B. Saha, S. Frégonese, S. R. Panda, A. Chakravorty, D. Céli, and T. Zimmer. Collector-substrate modeling of sige hbts up to thz range. In *2019 IEEE BiCMOS and Compound semiconductor Integrated Circuits and Technology Symposium (BCICTS)*, pages 1–4, November 2019. doi: 10.1109/BCICTS45179.2019.8972745. 62

[101] R. Sakamaki and M. Horibe. Uncertainty analysis method including influence of probe alignment on on-wafer calibration process. *IEEE Transactions on Instrumentation and Measurement*, 68(6):1748–1755, March 2019. doi: 10.1109/TIM.2019.2907733. 104

[102] F. J. Schmückle, T. Probst, U. Arz, G. N. Phung, R. Doerner, and W. Heinrich. Mutual interference in calibration line configurations. In *2017 89th ARFTG Microwave Measurement Conference (ARFTG)*, pages 1–4, June 2017. doi: 10.1109/ARFTG.2017.8000823. 81

[103] M. Schröter and A. Chakravorty. *Compact Hierarchical Bipolar Transistor Modeling With HiCUM*. WORLD SCIENTIFIC, 2010. doi: 10.1142/7257. URL https://www.worldscientific.com/doi/abs/10.1142/7257. 62

[104] Rohde & Schwarz. R&S ZVA vector network analyzers specifications, 2016. URL https://scdn.rohde-schwarz.com/ur/pws/dl_downloads/dl_common_library/dl_brochures_and_datasheets/pdf_1/ZVA_dat-sw_en_5213-5680-22_v1100.pdf. 33

[105] Rohde & Schwarz. Vector network analyzers, 2020. URL https://www.rohde-schwarz.com/us/products/test-and-measurement/network-analyzers/pg_overview_64043.html?rusprivacypolicy=0. 17

[106] Rohde & Schwarz. R&S ZCxxx millimeter-wave converters specifications, 2020. URL https://scdn.rohde-schwarz.com/ur/pws/dl_downloads/dl_common_library/dl_brochures_and_datasheets/pdf_1/ZCxxx_dat-sw_en_3607-1471-22_v1500.pdf. 33

[107] M. Seelmann-Eggebert, M. Ohlrogge, R. Weber, D. Peschel, H. Maßler, M. Riessle, A. Tessmann, A. Leuther, M. Schlechtweg, and O. Ambacher. On the Accurate Measurement and Calibration of S-Parameters for Millimeter Wavelengths and Beyond. *IEEE Transactions on Microwave Theory and Techniques*, 63(7):2335–2342, July 2015. doi: 10. 50

[108] K. Sengupta, T. Nagatsuma, and D. Mittleman. Terahertz integrated electronic and hybrid electronic–photonic systems. *Nature Electronics*, 1, 12 2018. doi: 10.1038/s41928-018-0173-2. 3, 8, 10

[109] Choon Beng Sia. Improving Wafer-Level S-parameters Measurement Accuracy and Stability with Probe-Tip Power Calibration up to 110 GHz for 5G Applications. page 4, October 2019. 32

[110] SURA. Terahertz applications symposium, 2009. http://www.sura.org/commercialization/terahertz.html. 3

[111] S. M. Sze. *High-Speed Semiconductor Devices*. J. Wiley, September 1990. ISBN 0471623075. URL http://www.ebook.de/de/product/4241998/high_speed_semiconductor_devices.html. 6, 7, 11

[112] Keysight Technologies. Agilent N5250A PNA millimeter-wave network analyzer 10 MHz to 110 GHz, 2004. URL https://www.brltest.com/pdf/pdf_analyzers/382.pdf. 32, 33

[113] Keysight Technologies. Keysight E8361A/C PNA network analyzer, 2014. URL https://literature.cdn.keysight.com/litweb/pdf/E8361-90007.pdf?id=451064. 32, 33

[114] Keysight Technologies. Keysight technologies N5260A millimeter head controller, 2017. URL http://literature.cdn.keysight.com/litweb/pdf/N5260-90001.pdf. 32

[115] Keysight Technologies. HICUM Model: Bipolar Transistor Model, 2019. URL https://edadocs.software.keysight.com/pages/viewpage.action?pageId=5923844. 63

[116] V. Teppati, A. Ferrero, and M. Sayed, editors. *Modern RF and Microwave Measurement Techniques*. The Cambridge RF and Microwave Engineering Series. Cambridge University Press, Cambridge, 2013. ISBN 978-1-107-03641-3. doi: 10.1017/CBO9781139567626. URL https://www.cambridge.org/core/books/modern-rf-and-microwave-measurement-techniques/8DC68F8105C26E37B65258FFC5DCDD47. 19

[117] L. F. Tiemeijer and R. J. Havens. A calibrated lumped-element de-embedding technique for on-wafer RF characterization of high-quality inductors and high-speed transistors. *IEEE Transactions on Electron Devices*, 50(3):822–829, june 2003. doi: 10.1109/TED.2003.811396. 32

[118] L. F. Tiemeijer, R. M. T. Pijper, J. A. van Steenwijk, and E. van der Heijden. A new 12-term open–short–load de-embedding method for accurate on-wafer characterization of rf mosfet structures. *IEEE Transactions on Microwave Theory and Techniques*, 58(2):419–433, February 2010. doi: 10.1109/TMTT.2009.2038453. 87

[119] T. Urbanec. Wideband electronic calibration set for sixport measurement systems. In *2009 IEEE International Conference on Microwaves, Communications, Antennas and Electronics Systems*, pages 1–4, November 2009. doi: 10.1109/COMCAS.2009.5385982. 29

[120] VDI. Vector network analyzer extenders, 2018. URL https://vadiodes.com/en/products/vector-network-analyzer-extension-modules. 32

[121] S. P. Voinigescu, E. Dacquay, V. Adinolfi, I. Sarkas, A. Balteanu, A. Tomkins, D. Celi, and P. Chevalier. Characterization and Modeling of an SiGe HBT Technology for Transceiver Applications in the 100–300-GHz Range. *IEEE Transactions on Microwave Theory and Techniques*, 60(12):4024–4034, December 2012. doi: 10.1109/TMTT.2012.2224368. URL http://ieeexplore.ieee.org/document/6355965/. 31

[122] V.-T. Vu, D. Céli, T. Zimmer, S. Fregonese, and P. Chevalier. Advanced si/sige hbt architecture for 28-nm fd-soi bicmos. 09 2016. doi: 10.1109/BCTM.2016.7738955. 14

[123] O. Wada and H. Hasegawa. *InP-Based Materials and Devices: Physics and Technology*. Wiley Series in Microwave and Optical Engineering. Wiley-Interscience, 1999. ISBN 978-0471181910. 7, 8, 11

[124] N. Waldhoff, C. Andrei, D. Gloria, S. Lepilliet, F. Danneville, and G. Dambrine. Improved characterization methology for mosfets up to 220 GHz. *IEEE Transactions on Microwave Theory and Techniques*, 57(5):1237–1243, April 2009. doi: 10.1109/TMTT.2009.2017359. 28, 87

[125] N. Waldhoff, C. Andrei, D. Gloria, S. Lepilliet, F. Danneville, and G. Dambrine. Improved characterization methology for mosfets up to 220 GHz. *IEEE Transactions on Microwave Theory and Techniques*, 57(5):1237–1243, 2009. doi: 10.1109/TMTT.2009.2017359. 87

[126] P.J. Wijnen, H.R. Claessen, and E.A. Wolsheimer. A new straightforward calibration and correction procedure for on wafer high frequency s-parameter measurements (45 MHz-18 GHz). *IEEE Proceedings of the Bipolar Circuits and Technology Meeting*, pages 70–73, January 1987. 30

[127] D. F. Williams, R. B. Marks, and A. Davidson. Comparison of on-wafer calibrations. *38th ARFTG Conference Digest*, pages 68–81, Dec 1991. doi: 10.1109/ARFTG.1991.324040. 65

[128] D. F. Williams, U. Arz, and H. Grabinski. Accurate characteristic impedance measurement on silicon. In *1998 IEEE MTT-S International Microwave Symposium Digest*, volume 3, pages 1917–1920 vol.3, June 1998. doi: 10. 65

[129] D. F. Williams, U. Arz, and H. Grabinski. Characteristic-impedance measurement error on lossy substrates. *IEEE Microwave and Wireless Components Letters*, 11(7):299–301, July 2001. doi: 10.1109/7260.933777. 65

[130] D. F. Williams, P. Corson, J. Sharma, H. Krishnaswamy, W. Tai, Z. George, D. Ricketts, P. Watson, E. Dacquay, and S. P. Voinigescu. Calibration-Kit Design for Millimeter-Wave Silicon Integrated Circuits. *IEEE Transactions on Microwave Theory and Techniques*, 61(7):2685–2694, July 2013. doi: 10.1109/TMTT.2013.2265685. 26, 28, 53, 56, 65

[131] D. F. Williams, A. C. Young, and M. Urteaga. A prescription for sub-millimeter-wave transistor characterization. *IEEE Transactions on Terahertz Science and Technology*, 3(4):433–439, April 2013. doi: 10.1109/TTHZ.2013.2255332. 42, 53, 56

[132] D. F. Williams, P. Corson, J. Sharma, H. Krishnaswamy, W. Tai, Z. George, D. S. Ricketts, P. M. Watson, E. Dacquay, and S. P. Voinigescu. Calibrations for millimeter-wave silicon transistor characterization. *IEEE Transactions on Microwave Theory and Techniques*, 62(3):658–668, January 2014. doi: 10.1109/TMTT.2014.2300839. 26, 27, 28, 29, 53, 71, 72

[133] D. F. Williams, F. Schmückle, R. Doerner, G. N. Phung, U. Arz, and W. Heinrich. Crosstalk corrections for coplanar-waveguide scattering-parameter calibrations. *IEEE Transactions on Microwave Theory and Techniques*, 62(8):1748–1761, July 2014. doi: 10.1109/TMTT.2014.2331623. 54, 56

[134] D.F. Williams and R.B. Marks. Transmission line capacitance measurement. *IEEE Microwave and Guided Wave Letters*, 1(9):243–245, 1991. doi: 10.1109/75.84601. 66

[135] Dylan F. Williams and Roger B. Marks. Calibrating On-Wafer Probes to the Probe Tips. In *40th ARFTG Conference Digest*, volume 22, pages 136–143, December 1992. doi: 10.1109/ARFTG.1992.327008. 29

[136] J. Winkel. Extended charge-control model for bipolar transistors. *IEEE Transactions on Electron Devices*, 20(4):389–394, April 1973. doi: 10.1109/T-ED.1973.17660. 9

[137] O. Wohlgemuth, M. J. W. Rodwell, R. Reuter, J. Braunstein, and M. Schlechtweg. Active probes for network analysis within 70-230 ghz. *IEEE Transactions on Microwave Theory and Techniques*, 47(12):2591–2598, 1999. 27

[138] W. Xiaoyun, G. Niu, S. Sweeney, Q. Liang, X. Wang, and S. Taylor. A general 4-port solution for 110 GHz on-wafer transistor measurements with or without impedance standard substrate (ISS) calibration. *Electron Devices, IEEE Transactions on*, 54:2706 – 2714, 11 2007. doi: 10.1109/TED.2007.904362. 87

[139] L. Xie, C. M. Moore, M. E. Cyberey, S. Nadri, N. D. Sauber, M. F. Bauwens, A. W. Lichtenberger, N. S. Barker, and R. M. Weikle. Micromachined probes with integrated GaAs schottky diodes for on-wafer temperature sensing. In *2018 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6, May 2018. doi: 10.1109/I2MTC.2018.8409690. 27

[140] L. Xie, M. F. Bauwens, S. Nadri, M. E. Cyberey, A. Arsenovic, A. W. Lichtenberger, N. S. Barker, and R. M. Weikle. Electronic calibration of one-port networks at submillimeter wavelengths using schottky diodes as on-wafer standards. In *2019 93rd ARFTG Microwave Measurement Conference (ARFTG)*, pages 1–4, June 2019. doi: 10.1109/ARFTG.2019.8739175. 29

[141] C. Yadav, M. Deng, M. De Matos, S. Fregonese, and T. Zimmer. Importance of complete characterization setup on on-wafer TRL calibration in sub-THz range. In *2018 IEEE International Conference on Microelectronic Test Structures (ICMTS)*, pages 197–201, March 2018. doi: 10.1109/ICMTS.2018.8383798. 81

[142] C. Yadav, M. Deng, S. Fregonese, M. De Matos, B. Plano, and T. Zimmer. Impact of on-Silicon De-Embedding Test Structures and RF Probes Design in the Sub-THz Range. In *2018 48th European Microwave Conference (EuMC)*, pages 21–24, September 2018. doi: 10.23919/EuMC.2018.8541392. 50, 79

[143] C. Yadav, S. Fregonese, M. Deng, M. Cabbia, M. De Matos, M. Jaoul, and T. Zimmer. Analysis of test structure design induced variation in on si on-wafer trl calibration in sub-thz. In *2019 IEEE 32nd International Conference on Microelectronic Test Structures (ICMTS)*, pages 132–136, March 2019. doi: 10.1109/ICMTS.2019.8730962. 45

[144] C. Yadav, M. Deng, S. Fregonese, M. Cabbia, B. Plano, and T. Zimmer. Importance and Requirement of frequency band specific RF probes EM Models in sub-THz and THz Measurements up to 500 GHz. *IEEE Transactions on Terahertz Science and Technology*, pages 1–1, June 2020. doi: 10.1109/TTHZ.2020.3004517. 61, 78, 79

[145] K. Yau, I. Sarkas, A. Tomkins, P. Chevalier, and S. Voinigescu. On-wafer s-parameter deembedding of silicon active and passive devices up to 170 ghz. pages 600 – 603, 06 2010. doi: 10.1109/MWSYM.2010.5518218. 87

[146] K. Yau, E. Dacquay, I. Sarkas, and S. P. Voinigescu. Device and ic characterization above 100 ghz. *IEEE Microwave Magazine*, 13(1):30–54, January 2012. doi: 10.1109/MMM.2011.2173869. 22, 25, 26, 28, 29

[147] K.H.K. Yau, A. M. Mangan, P. Chevalier, P. Schvan, and S. P. Voinigescu. A transmission-line based technique for de-embedding noise parameters. *2007 IEEE International Conference on Microelectronic Test Structures*, pages 237–242, March 2007. doi: 10.1109/ICMTS.2007.374491. 87

[148] Q. Yu, M. F. Bauwens, C. Zhang, A. W. Lichtenberger, R. M. Weikle, and N. S. Barker. Improved micromachined terahertz on-wafer probe using integrated strain sensor. *IEEE Transactions on Microwave Theory and Techniques*, 61(12):4613–4620, November 2013. doi: 10.1109/TMTT.2013.2288602. 27

[149] C. Zhang, M. Bauwens, N. S. Barker, R. M. Weikle, and A. W. Lichtenberger. A W-band micromachined on-wafer probe with integrated balun for characterization of differential circuits. *IEEE Transactions on Microwave Theory and Techniques*, 64(5):1585–1593, March 2016. doi: 10.1109/TMTT.2016.2538760. 27

[150] C. Zhang, M. Bauwens, M. E. Cyberey, L. Xie, A. W. Lichtenberger, N. Scott Barker, and R. M. Weikle. A differential probe with integrated balun for on-wafer measurements in the WR-3.4 (220 – 330 GHz) waveguide band. In *2019 IEEE MTT-S International Microwave Symposium (IMS)*, pages 1269–1271, May 2019. doi: 10.1109/MWSYM.2019.8701058. 27

[151] T. Zimmer, F. Sebastien, A. Curutchet, M. Potéreau, and C. Raya. Dispositif de calibrage pour l'ajustement d'une mesure radiofréquence., June 2015. URL https://hal.archives-ouvertes.fr/hal-01721675. Patent no. FR 15/56033. 104

List of Publications

Journal Papers

- M. Cabbia, C. Yadav, M. Deng, S. Fregonese, M. De Matos and T. Zimmer, "Silicon Test Structures Design for Sub-THz and THz Measurements", in *IEEE Transactions on Electron Devices*, vol. 67, no. 12, pp. 5639-5645, Dec. 2020, doi: 10.1109/TED.2020.3031575.
- M. Cabbia, S. Fregonese, M. Deng, A. Curutchet, C. Yadav, D. Céli, M. De Matos and T. Zimmer, "Meander-Type Lines: An Innovative Design for On-Wafer TRL Calibration for mmW and sub-mmW Frequencies Measurements", under review in *IEEE Transactions on Terahertz Science and Technology*.

Conference Papers (with Oral Presentations)

- M. Cabbia, M. Deng, S. Fregonese, M. D. Matos, D. Céli and T. Zimmer, "*In-Situ* Calibration and De-Embedding Test Structure Design for SiGe HBT On-Wafer Characterization up to 500 GHz", *94th ARFTG Microwave Measurement Symposium (ARFTG)*, San Antonio, TX, USA, 2020, pp. 1-4, doi: 10.1109/ARFTG47584.2020.9071733.
- M. Cabbia, M. Deng, S. Fregonese, C. Yadav, A. Curutchet, M. De Matos, D. Céli, and T. Zimmer, "Meander-Type Transmission Line Design for On-Wafer TRL Calibration up to 330 GHz", accepted for *European Microwave Conference 2020*.

Workshops

- M. Cabbia, M. Deng, C. Yadav, S. Fregonese, M. de Matos, and T. Zimmer, “Caractérisation RF de transistors bipolaires à hétérojonction SiGe jusqu’à 500 GHz”, *XIIIème colloque national du GDR SOC²,* Jun 2019, Montpellier, France.
- M. Cabbia, “RF Characterization of SiGe Heterojunction Bipolar Transistors in the Sub-THz Field ”, *Labo commun STMicroelectronics,* Sept 2019, Crolles, France.
- M. Cabbia, “In-Situ and Meander Test Structures: Two Innovative Calibration and De-Embedding Test Structures’ Designs for SiGe HBT On-Wafer Characterization up to 500 GHz”, *Labo commun STMicroelectronics,* Dec 2019, *held by remote*.

List of Publications

Journal Papers

- C. Yadav, M. Deng, S. Fregonese, M. Cabbia, M. De Matos, B. Plano and T. Zimmer, "Importance and Requirement of Frequency Band Specific RF Probes EM Models in Sub-THz and THz Measurements up to 500 GHz," in *IEEE Transactions on Terahertz Science and Technology*, vol. 10, no. 5, pp. 558-563, Sept. 2020, doi: 10.1109/TTHZ.2020.3004517.
- S. Fregonese, M. Cabbia, C. Yadav, M. Deng, S. R. Panda, M. De Matos, D. Céli, A. Chakravorty, T. Zimmer, "Analysis of High-Frequency Measurement of Transistors Along With Electromagnetic and SPICE Co-simulation," in *IEEE Transactions on Electron Devices*, vol. 67, no. 11, pp. 4770-4776, Nov. 2020, doi: 10.1109/TED.2020.3022603.
- S. Fregonese, M. Deng, M. Cabbia, C. Yadav, M. De Matos and T. Zimmer, "THz Characterization and Modeling of SiGe HBTs: Review (Invited)," in *IEEE Journal of the Electron Devices Society*, vol. 8, pp. 1363-1372, 2020, doi: 10.1109/JEDS.2020.3036135.
- C. Yadav, M. Cabbia, S. Fregonese, M. Deng, M. D. Matos, B. Plano, and T. Zimmer, “Guideline of Test Structures placement for on-Wafer Measurement of Electron Devices in sub-THz” in *IEEE Transactions on Components, Packaging and Manufacturing Technology* (submitted).
- C. Yadav, M. Deng, S. Fregonese, M. Cabbia, M. D. Matos, and T. Zimmer, “Investigation of Degradation in on-Si On-wafer TRL Calibration in sub-THz”, in *IEEE Transactions on Semiconductor Manufacturing* (submitted).

Conference Papers

- C. Yadav, S. Fregonese, M. Deng, M. Cabbia, M. De Matos, M. Jaoul, and T. Zimmer, "Analysis of Test Structure Design Induced Variation in on Si On-wafer TRL Calibration in sub-THz," *2019 IEEE 32nd International Conference on Microelectronic Test Structures (ICMTS)*, Kita-Kyushu City, Fukuoka, Japan, 2019, pp. 132-136, doi: 10.1109/ICMTS.2019.8730962.
- C. Yadav, S. Fregonese, M. Deng, M. Cabbia, M. De Matos and T. Zimmer, "On the Variation in Short-Open De-embedded S-parameter Measurement of SiGe HBT upto 500 GHz," *2019 12th German Microwave Conference (GeMiC)*, Stuttgart, Germany, 2019, pp. 264-267, doi: 10.23919/GEMIC.2019.8698153.

Workshops

- S. Fregonese, M. Deng, M. Cabbia, C. Yadav, S. R. Panda, T. Zimmer, “On wafer small signal characterization beyond 100 GHz for compact model assessment” in *European Microwave Week Workshop*, 2019.

Appendix A

Two-Port Representations

There exist different types of two-port network representations into matrix:

- Z-parameters (impedance)
- Y-parameters (admittance)
- h-parameters (hybrid)
- ABCD-parameters (chain)
- S-parameters (scattering)
- T-parameters (chain transfer)

S- and T-parameters are dependent on the source and load impedances.

If in cascade of two-port networks, a two-port matrix represent *output* voltage and current of one network to the *input* voltage and current of the following network then representation type is called transmission-type matrix. ABCD and T matrix are transmission-type matrices.

In a cascade two-port linear network, the waves at the output of first network is same as the waves at the input of second network (Fig. A.1).

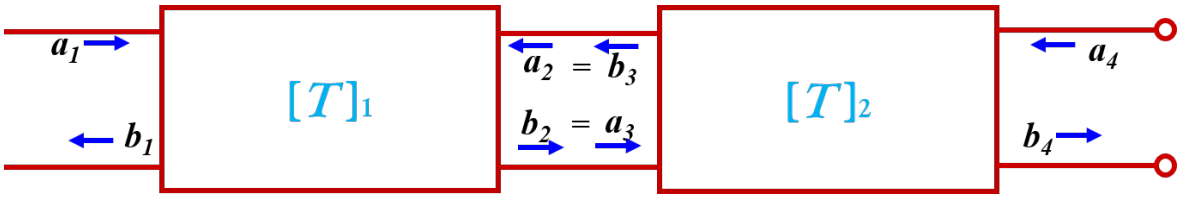


Figure A.1: Linear two-port networks connected in cascade.

Since $a_3 = b_2$ and $a_2 = b_3$, the relationship between the a_1, b_1 waves and a_4, b_4 waves can be obtained by simply multiplying the individual transmission T matrix. In general,

$$\begin{bmatrix} a_1 \\ b_1 \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} b_2 \\ a_2 \end{bmatrix} \tag{A.1}$$

with the following formulas to convert S-parameters to T-parameters:

$$\begin{aligned} T_{11} &= \frac{1}{S_{21}} & T_{12} &= -\frac{S_{22}}{S_{21}} \\ T_{21} &= \frac{S_{11}}{S_{21}} & T_{22} &= \frac{S_{12}S_{21} - S_{11}S_{22}}{S_{21}} \end{aligned} \tag{A.2}$$

and vice-versa, T-parameters to S-parameters:

$$\begin{aligned} S_{11} &= \frac{T_{21}}{T_{11}} & S_{12} &= \frac{T_{11}T_{22} - T_{12}T_{21}}{T_{11}} \\ S_{21} &= \frac{1}{T_{11}} & S_{22} &= -\frac{T_{21}}{T_{11}} \end{aligned} \tag{A.3}$$

Appendix B

TRL Calibration Algorithm

B.1 Computation of the Error Terms

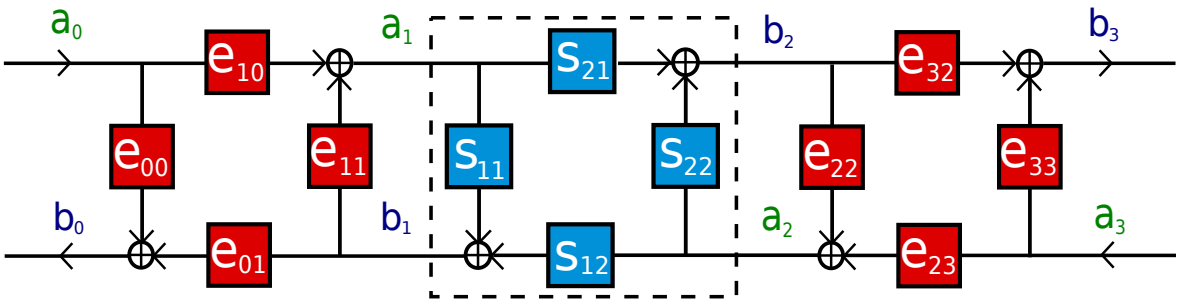


Figure B.1: 8-term error model for TRL calibration.

If we consider the 8-term error model (Fig. B.1), we can define the T-parameters for the error adapters at the left of the DUT (X) and on the right of the DUT (Y) as (from Eq. A.2):

$$[\mathrm{T_X}] = \begin{bmatrix} \mathrm{X_{11}} & \mathrm{X_{12}} \\ \mathrm{X_{21}} & \mathrm{X_{22}} \end{bmatrix} = \begin{bmatrix} \frac{e_{10}e_{01} - e_{00}e_{11}}{e_{10}} & \frac{e_{00}}{e_{10}} \\ -\frac{e_{11}}{e_{10}} & \frac{1}{e_{10}} \end{bmatrix} \tag{B.1}$$

and

$$[\mathrm{T_Y}] = \begin{bmatrix} \mathrm{Y_{11}} & \mathrm{Y_{12}} \\ \mathrm{Y_{21}} & \mathrm{Y_{22}} \end{bmatrix} = \begin{bmatrix} \frac{e_{32}e_{23} - e_{22}e_{33}}{e_{32}} & \frac{e_{22}}{e_{32}} \\ -\frac{e_{33}}{e_{32}} & \frac{1}{e_{32}} \end{bmatrix} \tag{B.2}$$

Any measured DUT can be expressed as:

$$[\mathrm{T_M}] = [\mathrm{T_X}] \cdot [\mathrm{T_{DUT}}] \cdot [\mathrm{T_Y}] \tag{B.3}$$

We also know that the T-parameters of a general ideal transmission line of length l are expressed as:

$$[\mathrm{T_L}] = \begin{bmatrix} e^{-\gamma l} & 0 \\ 0 & e^{\gamma l} \end{bmatrix} \tag{B.4}$$

so for a zero length thru, it yields:

$$[\mathrm{T_T}] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{B.5}$$

We begin the TRL calibration.

Thru Measuring the zero-length thru, we have (from Eq. B.3, B.5):

$$[\mathrm{T_{MT}}] = [\mathrm{T_X}] \cdot [\mathrm{T_Y}] \tag{B.6}$$

which gives:

$$[\mathrm{T_Y}] = [\mathrm{T_X}]^{-1} \cdot [\mathrm{T_{MT}}] \tag{B.7}$$

We also have from Eq. 2.10, measuring the reflection coefficient at port 1:

$$\Gamma_{\mathrm{M1}} = e_{00} + \frac{e_{10}e_{01}\,e_{22}}{1 - e_{11}e_{22}} \tag{B.8}$$

Also, measuring both of the transmission coefficients:

$$\mathrm{S_{21M}} = e_{10}e_{32}\,\frac{1}{1 - e_{11}e_{22}} \tag{B.9}$$

$$\mathrm{S_{12M}} = e_{23}e_{01}\,\frac{1}{1 - e_{11}e_{22}} \tag{B.10}$$

Line Now, measuring the line, we have (from Eq. B.3, B.4, B.7):

$$\begin{aligned} [\mathrm{T_{ML}}] &= [\mathrm{T_X}] \cdot [\mathrm{T_L}] \cdot [\mathrm{T_X}]^{-1} \cdot [\mathrm{T_{MT}}] \\ [\mathrm{T_{ML}}] \cdot [\mathrm{T_{MT}}]^{-1} &= [\mathrm{T_X}] \cdot [\mathrm{T_L}] \cdot [\mathrm{T_X}]^{-1} \\ [\mathrm{M}] &= [\mathrm{T_X}] \cdot [\mathrm{T_L}] \cdot [\mathrm{T_X}]^{-1} \end{aligned} \tag{B.11}$$

where we defined:

$$[\mathrm{M}] = [\mathrm{T_{ML}}] \cdot [\mathrm{T_{MT}}]^{-1} \tag{B.12}$$

By expanding:

$$\begin{aligned} m_{21}\left(\frac{\mathrm{X_{11}}}{\mathrm{X_{21}}}\right)^2 + (m_{22} - m_{11})\left(\frac{\mathrm{X_{11}}}{\mathrm{X_{21}}}\right) - m_{12} = 0 \\ m_{21}\left(\frac{\mathrm{X_{12}}}{\mathrm{X_{22}}}\right)^2 + (m_{22} - m_{11})\left(\frac{\mathrm{X_{12}}}{\mathrm{X_{22}}}\right) - m_{12} = 0 \end{aligned} \tag{B.13}$$

Both equations are in quadratic form and have same coefficients which leads to same (double) solution: a and b. This allows to find from Eq. B.1:

$$e_{00} = b = \frac{\mathrm{X_{12}}}{\mathrm{X_{22}}}; \quad \frac{e_{10}e_{01}}{e_{11}} = a - b = \frac{\mathrm{X_{11}}}{\mathrm{X_{21}}} - \frac{\mathrm{X_{12}}}{\mathrm{X_{22}}} \tag{B.14}$$

By the same procedure on $[\mathrm{T_Y}]$, we find from Eq. B.2:

$$e_{33} = -d = -\frac{\mathrm{Y_{12}}}{\mathrm{Y_{22}}}; \quad \frac{e_{23}e_{32}}{e_{22}} = c - d = \frac{\mathrm{Y_{11}}}{\mathrm{Y_{21}}} + \frac{\mathrm{Y_{12}}}{\mathrm{Y_{22}}} \tag{B.15}$$

Reflect We now measure the reflect, namely we connect at each port a termination with reflection coefficient Γ_{R}. From Eq. 2.10, we know that:

$$\Gamma_{\mathrm{MX}} = e_{00} + \frac{e_{10}e_{01}\,\Gamma_{\mathrm{R}}}{1 - e_{11}\,\Gamma_{\mathrm{R}}} \tag{B.16}$$

and

$$\Gamma_{\mathrm{MY}} = e_{33} + \frac{e_{23}e_{32}\,\Gamma_{\mathrm{R}}}{1 - e_{22}\,\Gamma_{\mathrm{R}}} \tag{B.17}$$

Gamma Moreover, by expanding and rearranging Eq. B.11, we get the propagation constant of the used line, γ from:

$$e^{2\gamma l} = \frac{bm_{21} + m_{22}}{\frac{1}{a}m_{12} + m_{11}} = \frac{m_{11} + m_{22} \pm \mathrm{R}}{m_{11} + m_{22} \mp \mathrm{R}} \Longrightarrow \gamma l = \frac{1}{2} \ln\left(\frac{m_{11} + m_{22} \pm \mathrm{R}}{m_{11} + m_{22} \mp \mathrm{R}}\right) \tag{B.18}$$

where we defined:

$$\mathrm{R} = \sqrt{(m_{11} - m_{22})^2 + 4m_{21}m_{12}} \tag{B.19}$$

Summing up By rearranging the previous equations (Eq. B.8, B.9, B.10, B.14, B.15, B.16, B.17) we obtain the final values of the error terms, fully dependent from measured results, shown in Table B.1.

$e_{00} = b$	$e_{33} = -d$
$e_{22} = \frac{1}{e_{11}}\left(\frac{b-\Gamma_{\mathrm{M1}}}{a-\Gamma_{\mathrm{M1}}}\right)$	$e_{11} = \pm\sqrt{\left(\frac{b-\Gamma_{\mathrm{MX}}}{a-\Gamma_{\mathrm{MX}}}\right)\left(\frac{c+\Gamma_{\mathrm{MY}}}{d+\Gamma_{\mathrm{MY}}}\right)\left(\frac{b-\Gamma_{\mathrm{M1}}}{a-\Gamma_{\mathrm{M1}}}\right)}$
$e_{10}e_{01} = (b-a)e_{11}$	$e_{23}e_{32} = (c-d)e_{22}$
$e_{10}e_{32} = \mathrm{S}_{21\mathrm{M}}(1 - e_{11}e_{22})$	$e_{23}e_{01} = \mathrm{S}_{12\mathrm{M}}(1 - e_{11}e_{22})$

Table B.1: Error terms.

We have successfully calculated the matrices $[\mathrm{T_X}]$ and $[\mathrm{T_Y}]$.

B.2 Calibration with Non-Zero Length Thru

The results found for the case of a zero thru can be converted to the more realistic case of a non-zero thru. In this situation, Eq. B.5 becomes:

$$[\mathrm{T_T}] = \begin{bmatrix} e^{-\gamma l_\mathrm{T}} & 0 \\ 0 & e^{\gamma l_\mathrm{T}} \end{bmatrix} \tag{B.20}$$

with $l_\mathrm{T} \neq 0$ the length of the thru. We therefore rename l_L the length of the line. Eq. B.12 can therefore be expressed as:

$$[\mathrm{M}] = [\mathrm{T_{ML}}] \cdot [\mathrm{T_{MT}}]^{-1} = \begin{bmatrix} e^{-\gamma(l_\mathrm{L} - l_\mathrm{T})} & 0 \\ 0 & e^{\gamma(l_\mathrm{L} - l_\mathrm{T})} \end{bmatrix} \tag{B.21}$$

The error terms solved for zero length thru can be used for the non-zero length thru, simply by substituting in formulas γl with $\gamma(l_\mathrm{L} - l_\mathrm{T})$.

Note that the reference plane is still at the center of the thru here. In order to change its position, we have to shift the reference plane through a matrix transformation:

$$\left[\mathrm{T'_{DUT}}\right] = [\mathrm{T}_h]^{-1} \cdot [\mathrm{T_{DUT}}] \cdot [\mathrm{T}_h]^{-1} \tag{B.22}$$

where $\left[\mathrm{T'_{DUT}}\right]$ is the transformed DUT T-matrix and $[\mathrm{T}_h]$ is the matrix defining half of the thru line, since we established that the reference plane should be moved to the edge of the thru. $[\mathrm{T}_h]$ is defined:

$$[\mathrm{T_T}] = \begin{bmatrix} e^{-\gamma l_\mathrm{T}/2} & 0 \\ 0 & e^{\gamma l_\mathrm{T}/2} \end{bmatrix} \tag{B.23}$$

Finally, by measuring the desired DUT we get $[\mathrm{T_M}]$ and reversing Eq. B.3, we obtain the actual T-parameters of the DUT:

$$[\mathrm{T_{DUT}}] = \left[[\mathrm{T_X}] \cdot [\mathrm{T}_h]^{-1}\right]^{-1} \cdot [\mathrm{T_M}] \cdot \left[[\mathrm{T}_h]^{-1} \cdot [\mathrm{T_Y}]\right]^{-1} \tag{B.24}$$

and through the formulas of Eq. A.3, the S matrix $[S_{DUT}]$.

Appendix C

Electric Quantities of a Line

Here is the definition of the parameter of a line by the Maxwell's equations [84]:

$$\begin{aligned} \mathrm{L} &= \frac{\mu}{|\mathrm{I}_o|^2}\int_{\mathrm{S}} \bar{\mathrm{H}}\cdot\bar{\mathrm{H}}^* d\mathrm{S}; \qquad \mathrm{C} = \frac{\varepsilon}{|\mathrm{V}_o|^2}\int_{\mathrm{S}} \bar{\mathrm{E}}\cdot\bar{\mathrm{E}}^* d\mathrm{S} \\ \mathrm{R} &= \frac{\mathrm{R}_s}{|\mathrm{I}_o|^2}\int_{\mathrm{C}_1+\mathrm{C}_2} \bar{\mathrm{H}}\cdot\bar{\mathrm{H}}^* dl; \qquad \mathrm{G} = \frac{\omega\varepsilon''}{|\mathrm{V}_o|^2}\int_{\mathrm{S}} \bar{\mathrm{E}}\cdot\bar{\mathrm{E}}^* d\mathrm{S} \end{aligned} \tag{C.1}$$

where: $|\mathrm{V}_o|$, $|\mathrm{I}_o|$ are the input voltage and current magnitudes; S is the cross sectional surface of the line; $\mathrm{C}_1+\mathrm{C}_2$ is the integration path on the metal boundaries (strip and ground); μ is the permeability; $\varepsilon = \varepsilon' - j\varepsilon''$ is the complex effective permittivity; $\mathrm{R}_s = \sqrt{\frac{\omega\mu}{2\sigma}}$ is the surface resistivity of the conductor, with σ, the conductivity of copper, and $\omega = 2\pi f$.

Appendix D

Simulations on Run 3

D.1 Shifted-Pads

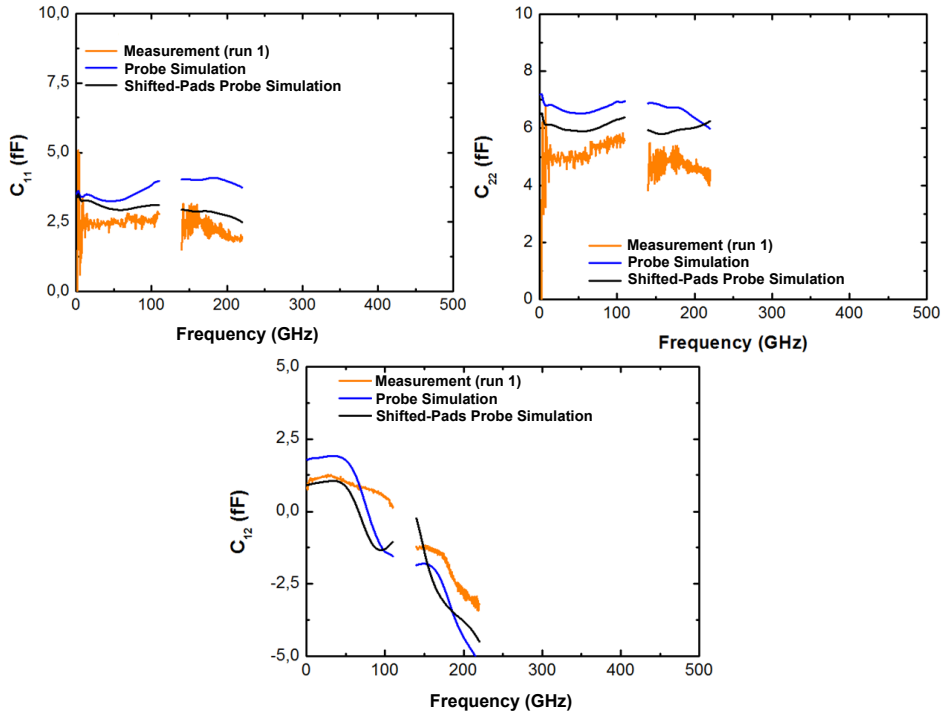
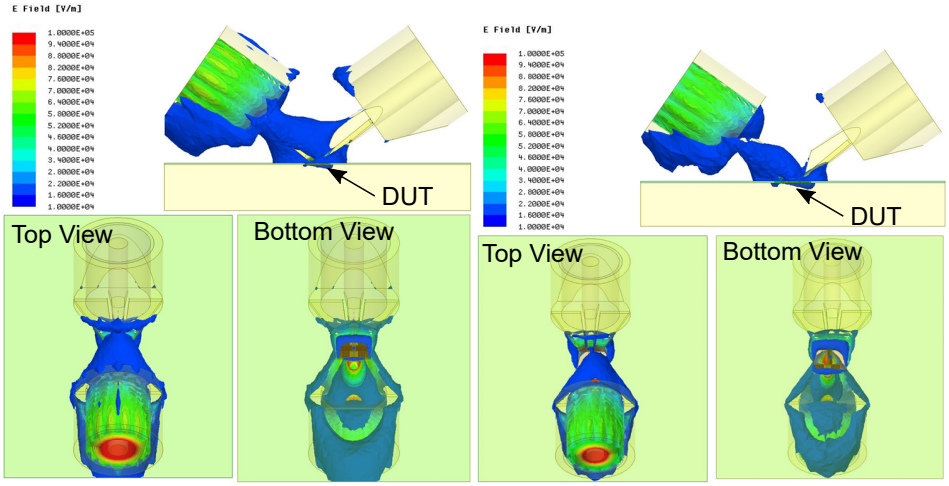


Figure D.1: Run 1 measurement of C-O, with the associated probe simulations (with classic TRL standards and shifted-pads standards). The shifted-pads standards have a prototype implementation by run 1 layout.



(a) Run 1 with standard TRL (b) Run 1 with shifted-pads TRL

Figure D.2: Comparison of E-field distributions at 110 GHz.

D.2 M6 TRL

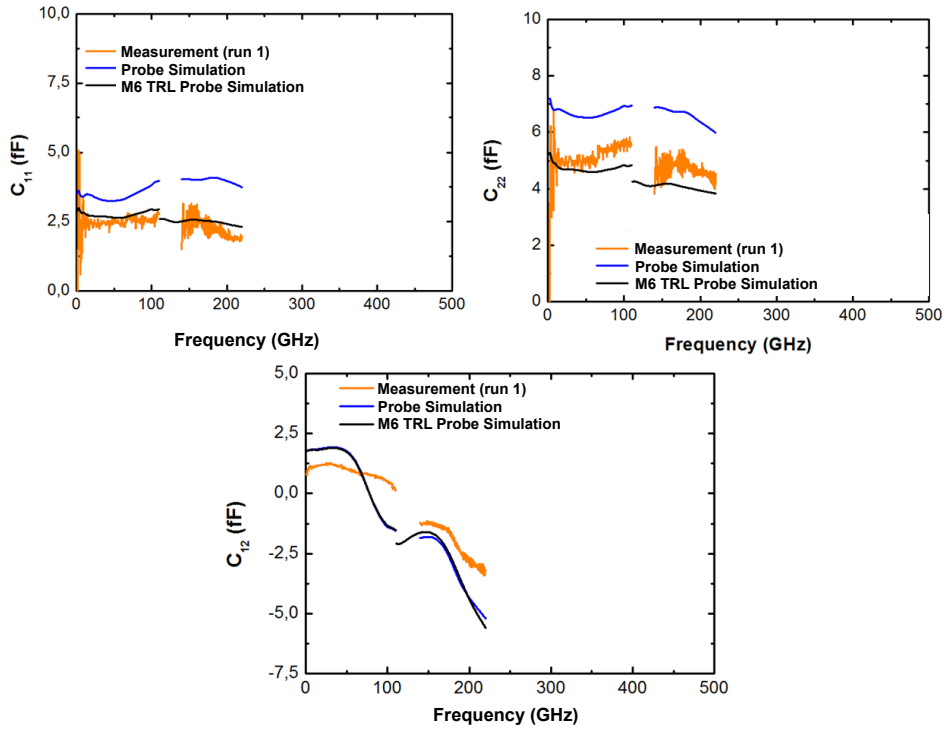
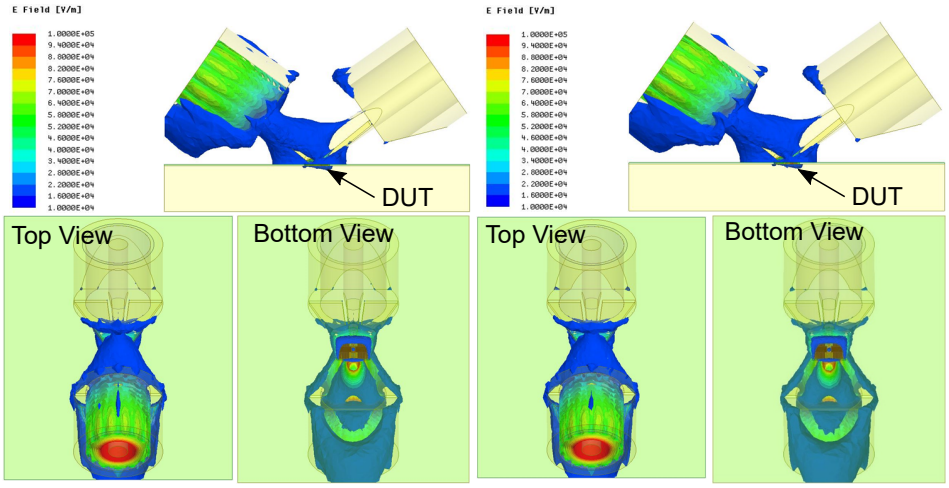


Figure D.3: Run 1 measurement of C-O, with the associated probe simulations (with classic TRL standards and M6 standards). The M6 standards have a prototype implementation by run 1 layout.



(a) Run 1 with standard TRL (b) Run 1 with M6 TRL

Figure D.4: Comparison of E-field distributions at 110 GHz.