Design and Modeling of 60-GHz CMOS Integrated Circuits

Chinh Huy Doan

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2009-134
http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-134.html

October 5, 2009

Design and Modeling of 60-GHz CMOS Integrated Circuits

by

Chinh Huy Doan

B.S. (California Institute of Technology) 1997
M.S. (University of California, Berkeley) 2000

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering-Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Robert W. Brodersen, Chair
Professor Ali M. Niknejad
Professor Paul K. Wright

Fall 2009

The dissertation of Chinh Huy Doan is approved:

Chair	Date
	Date
	Date

University of California, Berkeley

Fall 2009

Design and Modeling of 60-GHz CMOS Integrated Circuits

by Chinh Huy Doan

Abstract

Design and Modeling of 60-GHz CMOS Integrated Circuits

by

Chinh Huy Doan

Doctor of Philosophy in Engineering-Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Robert W. Brodersen, Chair

As the number of devices supporting high-quality digital multimedia continues to increase, there is a strong desire to transfer the data quickly and conveniently. Wireless transmission offers ease of setup, flexibility of placement, and avoids the need for unsightly and expensive cables. However, today's commercially available wireless systems are incapable of the multi-gigabit/sec data-rates required for this application. The frequency spectrum around 60-GHz is ideally suited for high-speed wireless data transfer, but it has yet to be widely used for consumer applications due to its high implementation costs.

Complementary metal-oxide-semiconductor (CMOS) processing is the lowest cost semiconductor technology. As a consequence of the ever-shrinking transistor dimensions, the high-frequency performance of state-of-the-art CMOS technology is improving, and it is becoming an attractive alternative to the expensive compound semiconductor technologies traditionally needed for 60-GHz transceivers. If a 60-GHz CMOS radio can be implemented, this would open up new opportunities for the ubiquitous use of this spectrum for consumer applications.

This dissertation explores the challenges involved in designing 60-GHz CMOS circuits. First, the key parasitic components that limit the high-frequency performance of CMOS transistors are identified, and an optimal layout to minimize these parasitic

elements is proposed. Second, transistor and passive models are investigated that can provide highly-accurate prediction of the device characteristics up to millimeter-wave (mm-wave) frequencies. This avoids the need to design with unnecessary margin due to modeling uncertainties. Following some basic guidelines, the models can result in simple extensions of commonly-used device models and are verified to be accurate up to 65 GHz. Accurate mm-wave measurements of the devices are necessary in order to extract good models, but the low resistivity of the CMOS substrate presents unique challenges. Different measurement and de-embedding methodologies are evaluated, and an approach to extract the small-signal and noise characteristics of the devices is presented and validated. Finally, mm-wave amplifiers and filters are fabricated in a 130-nm bulk digital CMOS technology to demonstrate the effectiveness of the device design and modeling methodology. This results in the first-reported 60-GHz CMOS amplifier and establishes the potential of using standard CMOS for fully-integrated 60-GHz transceivers.

Robert W. Brodersen, Chair

To Mom and Dad
for their enduring love and support.

Contents

Figures

Tables

Acknowledgment

The completion of a doctoral thesis is not the sole work of a single individual, but requires the help and support of many others along the way. It has been quite a long journey since my arrival at Berkeley in 1997, and I preemptively apologize to anybody that has contributed to my growth and development over the years as a graduate student researcher that I inadvertently omit in this acknowledgment. It is not out of a lack of appreciation but out of brevity to avoid any further delay of this thesis submission!

Foremost, I am deeply indebted and hold great respect for my research advisor Professor Robert Brodersen. You challenged me with a research project that nobody thought possible, and provided the resources and freedom to allow me to pursue the avenue that I thought best. Only with your foresight of 60-GHz CMOS being an important area of research was this project as successful as it was. I also want to thank Professor Ali Niknejad for serving on my dissertation and qualifying exam committee and who is carrying the torch at the Berkeley Wireless Research Center (BWRC) for 60-GHz CMOS circuits research. I have enjoyed multiple brainstorming discussions together and I am sure you will continue to make many great strides to advance this nascent field. I would also like to thank Professor Paul Wright for reviewing this dissertation and being on my qualifying exam committee and Professor Paul Gray for chairing my qualifying exam.

My close friend and colleague, Sohrab Emami, who has been my partner in crime for the past 8 years, directly contributed to the scope and quality of this work. Your dedication and perseverance to put together the prob station setup, *measure*, and extract excellent device models using IC-CRAP were instrumental to the success of this project. I hope we continue to share the same level of success together in the future. You also have the distinction of being the only person I know who can truly challenge me for the title of King Procrastinator. *Boro baa man, Benny*!!

I give sincere thanks to the original members of the Baseband Analog and RF (BARF) group—Dennis Yee, David Sobel, Brian Limketkai, Johan Vanderhaegen, Ian O'Donnell, Sayf Alalusi, and Sam Blackman—of which I am honored to have been a member. It was a very talented group of engineers with diverse backgrounds from which I learned many things, academically and otherwise. Thanks Dennis for being our fearless group leader and mentoring me on how to be truly passionate about circuit design. I also enjoyed many hours outside of the lab with you at Tilden Park, the pool hall, and the RSF. Thanks David for teaching me that there was always a fair way to do everything. Thanks Brian for providing countless hours of philosophical deliberations, teaching me to never share a basket of curly fries, and driving me to perfect this thesis template. Sorry that I won't have time to add a colophon! Thanks Johan for the Belgian beer and chocolates and for being available to discuss any problem, large or small, whether or not it was in your area of expertise. Thanks Ian for being a great cubicle-mate and always providing a joke to break the monotony of grad school life. Thanks Sayf for spotting me on the bench press. Thanks Sam for the stock tips and giving me a glimpse of what it takes to become an entrepreneur. I can honestly say that I wouldn't have been able to make it through grad school, sanity intact, without you guys!

I would like to thank Henry Jen for the many hours of discussing life, love, and the pursuit of the perfect golf swing. Thanks to Ada Poon, Andy Klein, and Chris Taylor for teaching me a thing or two about communications systems and looking out for me at the BWRC retreats. I also had the pleasure to interact with several senior graduate students who set the bar at Berkeley so high that it always motivated me to do better. In this group are Jeff Weldon, Chris Rudell, Tom Burd, Andy Abo, Keith Onodera, and Sehkar Narayanaswami. I would also like to thank Vace Shakoori for lending a cheerful smile and words of encouragement whenever I needed it (except on the ski slopes!).

When this project started, I decided to take a summer internship position at Agilent Technologies to learn microwave design directly from people who have been

doing it for decades. I would classify that experience as nearly ideal, and that summer has had a profound influence on the direction and outcome of this project. It provided the perspective I was looking for, dispelled some of my personal misconceptions about "traditional" microwave circuit design, and made me realize that these "microwave guys" are not so bad after all! I would particularly like to thank Tim Shirley, Dan Scherrer, Alex Cognata, Masaya Iwamoto, David Root, Kohei Fujii, and Hue Tran, for providing technical guidance and support and making my time at Agilent a great one!

I would like to thank several industrial members of the BWRC that helped me out through the years. Thanks to Bhusan Gupta, Isabelle Telliez, Ben Coates, and STMicroelectronics for helping me through so many tapeouts and chip fabrication. Thanks to Ken Kundert, Andrei Vladimirescu, Bob Frye, and Timo Kartaavi for technical discussions and help on simulations, packaging, and noise measurements.

I am also grateful to the administration staff at the BWRC and the EECS graduate office for helping me navigate through the bureaucracy of the UC system. In particular, Tom Boot, Elise Mills, and Ruth Gjerde were always of tremendous help and willing to go out of their way to fix whatever sticky situation I happened to get myself into.

There are those outside of Berkeley that I would also like to thank, namely my friends from Caltech. Thanks to Jason Hong for the many great misadventures that we've had and the hours analyzing why we overanalyze. As long as we don't talk about chemistry or engineering, we are best friends. I foresee a trip to Vegas in our near future! Gilbert Yoh and I were nicknamed Pinky and the Brain in college. Fortunately, I was the shorter one. Thanks for pushing me to hike through some amazing terrain and being a great travel companion whenever I needed time to just get away. Thanks also to James Lin (a.k.a. *Jimm-aaaay*), who was my roommate for 7 years at Berkeley. I've known you that long, and there is still nothing bad I can say about you.

Finally, I would like to express my sincere thanks to my family for their love and encouragement through the years. To my sisters—Lianna, Tran, Thuy—there are

times when being together is unbearable, but when I take a step back, I realize that you are only looking out for my best interests. While there are times when my actions may seem ungrateful, please know that deep down, I feel fortunate to have such caring sisters.

But most of all, I would like to thank Mom and Dad. You have continually made so many personal sacrifices in order to provide the kids with limitless opportunities and instilled in each of us the importance of having a good education. Although I will never be able to repay you for all that you have given to me over the years, this dissertation is but a small token of my appreciation. For that, I dedicate this work to you with all my love.

1

Introduction

1.1 Motivation

We live in an increasingly digital world. There is a proliferation of digital media in all aspects of our everyday lives—high-definition (HD) video content, digital cameras, portable music players, and increasingly larger compact storage for data and multimedia. With so much digital content and multiple source and sink devices, there is a strong desire to be able to transfer this data quickly and simply between the devices. Cables and high-speed signaling interfaces are the industry standard for high-speed data transfer because of their reliability and moderate cost. However, the web of unsightly cables places a constraint on the physical location of these components. If the data can be transmitted wirelessly, then this offers the ultimate in convenience of placement and avoids the cost of requiring many high-speed cables. In addition to just replacing the point-to-point functionality of a cable, if devices are connected wirelessly, then concepts familiar to wireless networks can be

applied to make the connections re-configurable in real-time and control information can be passed between any two devices within this network to allow new functions and usage models. Although there is a clear advantage for wireless high-speed connectivity, the bandwidth provided by today's wireless systems are far from what is required to support these multi-gigabit/sec applications. To understand what is needed to achieve wireless data-rates at over 10 times faster than what can be commercially achieved today, it is instructive to review the information theoretical result of Shannon's theorem for additive white Gaussian noise (AWGN) channels [1].

$$C = B\log_2\left(1+\frac{S}{N}\right) \quad \textbf{(1.1)}$$

where C is the channel capacity in bits/sec, B is the bandwidth of the channel in hertz, S is the total received power over the bandwidth in watts, and N is the total noise power over the bandwidth in watts. For a wireless channel to have the capacity to support multi-Gbps data transfer, the channel must have large bandwidth B and large allowable signal power S. Another practical consideration is that the wireless spectrum used should be *unlicensed* to allow for pervasive deployment without the need for expensive infrastructure costs and subscription. Fortunately, the 60-GHz spectrum exists that satisfies all of these requirements. The ability to exploit the 60-GHz band to wirelessly transfer data at multi-Gb/s at sufficiently low cost will have a tremendous impact on the way we think about connecting these multimedia devices in the future.

1.1.1 60-GHz unlicensed spectrum

The 60-GHz spectrum is ideally suited for high data-rate applications because of its unlicensed world-wide availability, large contiguous bandwidth, and high

allowable transmit power. In December 1995, 5 GHz of contiguous bandwidth was opened for unlicensed use by the Federal Communications Commision (FCC) at millimeter-wave (mm-wave) frequencies around 60 GHz [2]. This was subsequently extended to 7 GHz in the U.S. (57–64 GHz), providing 5 GHz of overlap with unlicensed spectrum in Japan (59–66 GHz) and other geographical regions. The significance of being unlicensed is that operators do not need to purchase an expensive license which allows widespread deployment and is the primary reason why today's wireless local area networks (WLANs) all operate in unlicensed bands around 2.4 GHz and 5 GHz.

Although there is a large amount of spectrum between 5 GHz and 60 GHz, nowhere below 60 GHz is there so much unlicensed spectrum world-wide. This is because of its unique propagation characteristics. Oxygen molecules in the atmosphere resonate at frequencies around 60 GHz which causes significant attenuation of the electromagnetic energy at distances beyond a few kilometers. In addition to the free-space path loss, which applies to all RF radiation and results in a square-law decrease in received signal power with distance, the path loss of 60-GHz radiation also increases exponentially with distance approximately at a rate of 15 dB/km. Long range transmission is not possible, but conversely, long range interference is not an issue. This natural isolation enables multiple systems to operate within close proximity of one another. For this reason, the regulatory agencies chose to set a high limit on the maximum allowable effective isotropic radiated power (EIRP) of 40 dBm (10 W) in the US and 57 dBm (500 W) in Japan. In order to deliver wireless data-rates more than an order of magnitude higher than what can be achieved today, it is necessary to exploit this special combination of large available bandwidth and high allowable transmit power.

1.1.2 60-GHz wireless standards

Applications that require high bandwidth wireless transmissions include streaming uncompressed high-definition video, increased capacity wireless local area networks (WLANs), short-range high data-rate wireless personal area networks (WPANs), and gigabit/sec point-to-point links. Recently, there have been strong advancements in standardization to realize wireless specifications at 60 GHz, including WirelessHD™ [3], IEEE 802.15.3c [4], 802.11ad [5], and Wireless Gigabit Alliance (WiGig™) [6], that vary in their scope and approach to address some of the different applications. Industry standards are necessary to ensure interoperability between devices manufactured by multiple vendors and coexistence between different systems sharing the same frequency spectrum.

The WirelessHD special interest group was formed in 2006 with members including Intel, LG Electronics, Matsushita Electric Industrial (Panasonic), NEC, Samsung Electronics, SiBEAM, Sony and Toshiba, to create the next generation wireless interface specification with primary focus on high-definition media streaming and transmission between fixed location and portable consumer electronics (CE) devices such as HDTVs, Blu-Ray and HD-DVD players, and HD camcorders. The key characteristics of the WirelessHD specification include [7]:

- High interoperability supported by major CE device manufacturers
- Uncompressed HD video, audio and data transmission, scalable to future high-definition A/V formats
- High-speed wireless, multi-gigabit/sec technology in the unlicensed 60-GHz band
- Smart antenna technology to overcome line-of-sight (LOS) constraints
- Secure communications with Hollywood-approved content protection
- Device control for simple operation of consumer electronics products

- Error protection, framing and timing control techniques for a quality consumer experience

Using the WirelessHD specification, seamless wireless interconnection and interoperation of a wide array of CE devices are possible. Integrated device control coupled with non-line-of-sight (NLOS) smart antenna techniques allows simple and intuitive operation in a variety of environments. Uncompressed audio and video allow the highest quality user experience, unburdened from the increased cost and latency and decreased quality of many compression technologies. A key differentiator between the WirelessHD specification is that the limited application target allows for an optimized specification incorporating wireless physical layer (PHY), media access control (MAC), audio/video control, and content protection.

In July 2003, the IEEE 802.15.3 working group for WPAN began investigating the use of the 7 GHz of unlicensed spectrum around 60 GHz as an alternate physical layer (PHY) to enable very high data-rate applications. This led to the formation of the IEEE 802.15.3 Task Group 3c (TG3c) in March 2005. This mm-wave WPAN will support applications requiring high data-rates of at least 1 Gbps such as high-speed internet access and streaming content download (video on demand, home theater, etc.). Optionally, data rates in excess of 2 Gbps will be supported to provide for simultaneous time-dependent applications such as real-time HDTV video streaming and wireless data bus for cable replacement.

More recently, the IEEE 802.11ad task group has proposed the use of the 60-GHz band for a very high throughput extension of the IEEE 802.11 family of WLAN connectivity. The WiGig Alliance was established by more than 15 technology leaders within the PC, CE, semiconductor, and handheld industries. Its mission is to unify the next generation of high-speed wireless products by creating a comprehensive specification that encourages the adoption and widespread use of 60-GHz wireless technology worldwide.

1.1.3 60-GHz and other wireless technologies

With so many different wireless standards vying to address similar markets, it still remains to be seen if one or multiple standards will emerge victorious. However, what is clear is that the industry needs wireless connectivity in excess of 1 Gbps, and current wireless technologies, such as 802.11n and ultra-wideband (UWB), cannot provide the necessary data-rates, as shown in Table 1.1.

Wireless standard	Total spectral availability (GHz)	Maximum allowed EIRP (dBm)	Maximum data rate (Mbps)	Channel bandwidth (MHz)	bps/Hz to achieve 4 Gbps
WirelessHD™	7.0	40	~4,000	~2,000	~2
802.11n	0.67	22–35	289	20	200
			600	40	100
UWB	1.5 – 7.5	−10	480	520	8

Table 1.1 Comparison of different high-speed wireless standards [7].

The completed WirelessHD specification is used as a concrete example of the capabilities at 60 GHz, but similar conclusions would also apply to the other 60-GHz standardization efforts described in Section 1.1.2. Supporting very high data rates requires either large bandwidth or very high spectral efficiency since data rate is calculated as follows:

$$R = B \times SE \tag{1.2}$$

where R is the raw data rate in bps, B is the occupied bandwidth in Hz, and SE is the spectral efficiency in bps/Hz.

Increasing bandwidth requires operation in a large frequency spectrum band, which for sufficient range and robustness must allow reasonable transmit power. While UWB has multiple gigahertz of available spectrum, its capabilities are severely limited because of the tight restrictions on the

allowable transmit power to avoid causing unwanted interference to other systems operating in the same frequency bands. The 60-GHz band does not suffer from such constraints. Furthermore, UWB is only approved for unlicensed use in limited geographical areas rather than being available worldwide.

For an 802.11n system to be able to deliver 4 Gbps data rates, this would require a spectral efficiency of 100 bps/Hz, even if its wider 40-MHz mode is used. Higher spectral efficiency through multi-input/multi-output (MIMO) techniques such as spatial multiplexing typically leads to higher cost and reduced range and robustness due to the need for multiple concurrent paths between the transmitter and receiver and hardware to separate the data traveling over these different paths. Currently, a spectral efficiency of 10-15 bps/Hz is targeted as the maximum spectral efficiency supported by 802.11n, and it is unlikely that the highest data rates will be achieved in most practical environments and implementations. Contrast this with a 60-GHz radio that uses 2 GHz of spectrum. This would only require a spectral efficiency of 2 bps/Hz, leading to simpler radio architectures and digital signal processing needed to recover the wireless data.

There is general industry acceptance that 60 GHz is the only technology available that can deliver more than 1 Gbps wirelessly. The suitability of 60 GHz for these applications was recognized when this spectrum was made available nearly 15 years ago [2]. What is the reason for the recent standardization efforts when this spectrum has been available for so long? A few critical developments to enable a low-cost 60-GHz solution that overcome some of the traditional limitations of this wireless technology have only recently been demonstrated and are discussed in the next section.

1.1.4 Why 60 GHz now?

The applications that are being pursued by these different 60-GHz standards are in-room, high-throughput wireless applications for consumer electronics. Two characteristics that are necessary for mass market acceptance are (1) ease of setup and use and (2) low implementation costs.

60-GHz coverage angle and antenna gain In order for the wireless device to have simple setup, both the *coverage angle* and *range* must be sufficient to sustain the high data-rates in typical indoor environments. It has traditionally been challenging to achieve both of these requirements *simultaneously* at 60 GHz. Examination of Friis propagation loss formula can explain why.

Assuming simple line-of-sight (LOS) free-space communication, the Friis propagation loss is given by

$$\frac{P_r}{P_t} = \frac{D_r D_t \lambda^2}{(4\pi R)^2} = \frac{D_r D_t c^2}{(4\pi R f)^2} \qquad \textbf{(1.3)}$$

where P_r is the received signal power, P_t is the transmitted signal power, D_r is the antenna directivity of the receiver, D_t is the antenna directivity of the transmitter, R is the distance, λ is the free-space wavelength, c is the speed of light, and f is the operating frequency.

For a simple antenna system that is static without beam-steering technology, if the coverage angle is assumed to remain constant, then D_r and D_t are fixed, and for a given distance, (1.3) simplifies to

$$\frac{P_r}{P_t} \propto \left(\frac{1}{f}\right)^2 \qquad \textbf{(1.4)}$$

Comparing a 60-GHz system with a 5-GHz system, (1.4) shows that the signal at 60 GHz experiences 21.6 dB higher path loss. Clearly, maintaining the same coverage angle using a fixed antenna system is not a good solution.

The reason that (1.4) degrades so drastically at higher frequencies is because when the directivity (i.e., coverage) is kept constant, the antenna dimensions are scaled inversely proportional to the frequency. This is shown by the relationship

$$D = \frac{4\pi A}{\lambda^2} = \frac{4\pi A f^2}{c^2} \tag{1.5}$$

where D is the antenna directivity, A is the antenna aperture area, λ is the free-space wavelength, c is the speed of light, and f is the operating frequency. A different approach is to constrain the form factor of the antenna, rather than the coverage, and compare the received signal power. If the antenna aperture A is held constant, then substituting (1.5) into (1.3), we find that

$$\frac{P_r}{P_t} \propto f^2 \tag{1.6}$$

For a fixed form factor, there is a 21.6 dB *increase* in the received signal level at 60 GHz because the antenna gain on both the transmitter and receiver is increased. A directive antenna pattern also improves the channel multipath profile; by limiting the spatial extent of the transmitting and receiving antenna patterns to the dominant transmission path, the delay spread and Rician *K*-factor of an indoor wireless channel can be significantly improved [8].

The antenna directivity is traditionally increased by using physically large antennas, such as a horn antenna or dielectric lens, to focus the radiation energy. However, fixed high-gain antennas are not acceptable for typical consumer electronics since they require precise manual alignment and only operate well for point-to-point communications in a line-of-sight channel with no intervening obstructions. A system employing antenna arrays with adaptive electronically-steerable beams can be used to simultaneously obtain high directivity and good radiation coverage.

Smart-antenna beam-steering technology An antenna array is composed of individual radiating elements that are arranged in space to produce a directional radiation pattern. In the typical case where the antenna array contains identical antenna elements, the total radiation pattern is

$$(Array\ pattern) = (Array\ factor) \times (element\ pattern) \qquad \textbf{(1.7)}$$

The array factor (*AF*) depends on the configuration, distance between antenna elements, and the amplitude and phase excitation of the elements. For a 2-D array with *N* elements, the *AF* can have a maximum value of *N* in any direction with proper phasing of the signals. The individual element pattern determines the coverage angle, while the configuration and excitation determines the direction of the beam. This is an important technique that can provide good coverage angle and large antenna gain simultaneously.

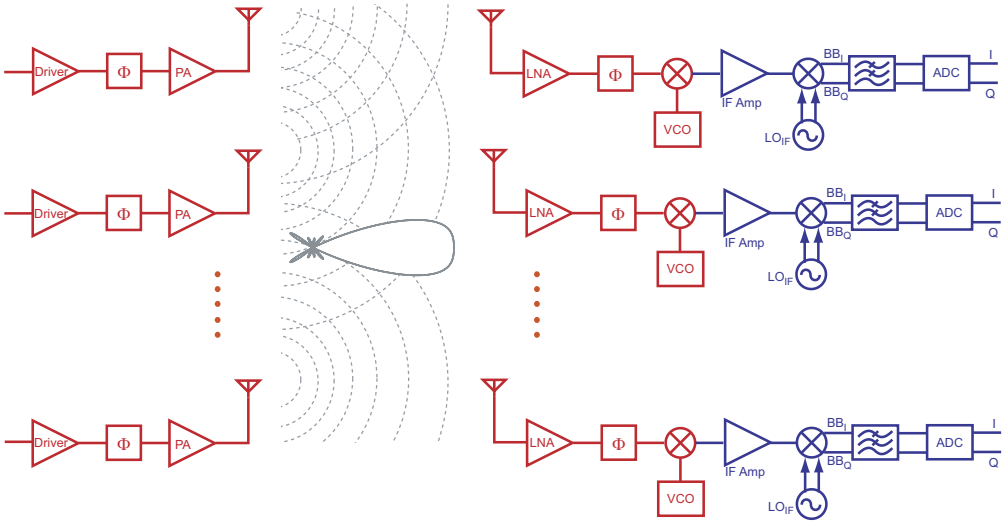


Figure 1.1 A generic multiple-antenna transceiver architecture employing beam steering.

Smart-antenna technology can be used to steer the directional beam electronically and adapt to a changing environment, automatically finding the optimal path in both direct LOS and NLOS paths that bounce off objects and walls. A generic adaptive beam-forming multiple antenna radio system is

shown in Fig. 1.1. This transceiver architecture depicts *N* independent transmit and receive chains. The main drawback to this multi-input/multi-output (MIMO) system is the high transceiver complexity and power consumption since there is little sharing of the hardware components. Measurements of the 60-GHz channel properties indicate that most of the received energy is contained in the specular path [8], so a full MIMO solution may not be necessary. A more efficient implementation would be to use a phased array that takes the identical RF signal and shifts the phase for each antenna to achieve beam steering. This drastically reduces hardware costs, as most of the transceiver can be shared with the addition of controllable phase shifters between the transceiver and antenna array.

In order to simultaneously achieve good coverage angle and large antenna gain using beam-steering technology, a multiple front-end transceiver needs to be implemented along with the digital signal processing to control the phase shifters. It is not possible to do this at low cost using compound semiconductor technologies. The recent advancement in silicon circuits is a promising technology to enable low-cost, highly-integrated 60-GHz transceivers.

60-GHz silicon circuits While compound semiconductor processing can achieve the performance at 60 GHz, it was never able to do so at the price points necessary for widespread consumer use. Digital-CMOS technology is the lowest cost option and allows the highest levels of integration to implement multiple transceivers along with the digital signal processing needed to design a robust wireless system all on a single chip. The advances over the past decade in CMOS technology and the immense research effort in RF CMOS circuit design techniques have made fully-integrated RF CMOS transceivers a reality. Inexpensive CMOS technologies have been used successfully to implement all the necessary RF functionality for the existing and emerging

wireless local/personal area network (WLAN/WPAN) standards, such as 802.11a/b/g/n, ultra-wideband (UWB), and Bluetooth. These systems all operate below 10 GHz. Moving to the largely unused spectrum at millimeter-wave (mm-wave) frequencies (> 30 GHz) will avoid the interference from other electronic devices operating in these lower bands and allow us to take advantage of the benefits of the 60-GHz frequency band.

Several recent developments have combined to enable CMOS circuit blocks to operate at ever-increasing frequencies. First, mm-wave CMOS circuits directly benefit from the higher speed of the scaled technology. 130-nm bulk-CMOS technology is capable of power gain at 60 GHz [9], and future bulk-CMOS processes at the 90-nm, 65-nm, and 45-nm nodes are expected to provide even more gain at lower power consumption. Secondly, improved circuit topologies and new design approaches to fully exploit the intrinsically faster devices have been introduced. Oscillators [10]–[12] were the first circuits demonstrated in CMOS to operate beyond 30 GHz, followed by CMOS amplifiers [13] and mixers [14], and culminating in fully-integrated 60-GHz CMOS front-ends [15].

With the possibility of a low-cost 60-GHz CMOS transceiver and the ability to integrate multiple front-ends with the digital signal processing needed for smart-antenna beam-steering technology, a 60-GHz transceiver finally appears possible to address the consumer electronics market need for a multi-Gbps wireless link.

1.2 Thesis Organization

This dissertation describes my initial efforts to use 130-nm digital-CMOS technology to exploit the 60-GHz band. The goal of this research is to investigate the fundamental frequency limitations of CMOS circuits, identify the challenges involved in designing robust mm-wave circuits in CMOS, and

propose and implement methods to overcome the barriers of using a technology at frequencies well beyond its original intended target to design circuits. Once these basic problems are understood and overcome, the cost advantages of using an inexpensive CMOS technology operating at 60-GHz opens up a wealth of new high data-rate wireless applications.

1.2.1 Fundamental frequency limits

The theoretical and practical high-frequency limitations of active devices are explored in Chapter 2. The figures-of-merit used to design analog and RF circuits are not directly relevant for determining the maximum operating frequency of a circuit in a given technology. Circuit theoretical results are derived for quantifying the mm-wave performance of CMOS, and the dominant parasitic elements are identified. Some parasitics are an intrinsic property of the manufacturing process, while others are extrinsic and layout-dependent. An optimized design of the core transistor is proposed in order to approach the intrinsic frequency limit of the CMOS process.

1.2.2 Millimeter-wave CMOS passives

As the size of passives needed for resonators and matching networks shrinks, becoming comparable to the interconnect wiring and transistor parasitics, the passives needed for mm-wave circuit designs must also be adapted. Transmission lines are familiar components in high-frequency circuit designs and can be used for matching networks, impedance transformation, and interconnect. Additionally, the well-defined ground return path ensures that no unwanted and unmodeled parasitic inductance is introduced into the circuit, making them well-suited for mm-wave circuits. The trade-offs in the design of CMOS transmission lines are analyzed and presented.

1.2.3 Device modeling

The lack of accurate CMOS active and passive device models at mm-wave frequencies is a key barrier for designing complex CMOS transceiver circuits at these frequencies, and a new approach must be considered. There is a vast gap in the device modeling methodology used for traditional mm-wave design and analog CMOS circuits. While the *S*-parameter models are accurate to very high frequencies, they also impose many constraints including being non-scalable and fixed bias. The transistor models used in analog CMOS circuit design are more flexible, but suffer from poor accuracy even at low microwave frequencies. This uncertainty forces the designer to either build a large margin into the circuit or require many iterations of the design to fine-tune the performance. A modeling methodology that blends the key characteristics of traditional mm-wave and analog CMOS modeling is needed to design optimized circuits near the frequency limits of the technology. A modeling methodology for both transistors and passives that results in simple, highly-accurate models up to 65 GHz is developed and validated in Chapter 2.

1.2.4 Microwave CMOS de-embedding

Device models are only as accurate as the measurements that they are derived from. There has been extensive research in the field of calibration and de-embedding for accurate on-wafer measurements. However, the methodology used for devices fabricated in CMOS has been mostly focused on frequencies up to only a few GHz. The different approaches to on-wafer measurements and de-embedding and their limitations are described. Metrics that are accurate to mm-wave frequencies are introduced that can help to evaluate the effectiveness of different de-embedding techniques. By fitting the models to the accurate measurements, circuits with predictable performance can be designed that are

operating at the edge of a given technology's capabilities.

1.2.5 Millimeter-wave circuit design

With all of the key building blocks optimized for mm-wave performance and accurate models that can predict their behavior to these frequencies, 40-GHz and 60-GHz wideband amplifiers and filters combining these core devices are designed and fabricated using a bulk 130-nm CMOS process to demonstrate the effectiveness of the approach. The 40-GHz amplifier attains 19-dB gain, output P_{1dB} = −0.9 dBm, IIP3 = −7.4 dBm, and consumes 36 mW. The 60-GHz amplifier achieves 12-dB gain, output P_{1dB} = +2.0 dBm, NF = 8.8 dB, and dissipates 54 mW. This represents the first reported amplifiers operating above 30 GHz, and confirms that CMOS can be used to design reliable circuits at mm-wave frequencies.

2

Design and Modeling of CMOS Devices

2.1 Introduction

When designing CMOS circuits for mm-wave operation, the use of a low-resistivity silicon substrate and the parasitic source, drain, and gate resistances are some key factors that degrade the electrical performance of the active and passive devices. In order to properly account for such limitations and minimize their impacts, new methodologies must be developed that involve optimized layouts, careful modeling, accurate measurement and de-embedding techniques, and a variety of simulation strategies. The methodology that has been developed for the design and modeling of both active and passive devices will be described along with experimental verification up to 65 GHz.

2.2 CMOS Technology

This section provides a brief comparison between a standard digital 130-nm

CMOS process and a dedicated microwave technology such as GaAs. Some of the key differences, which motivate the design choices and influence the modeling, are highlighted.

2.2.1 Front-end features

Two important disadvantages of a silicon metal-oxide-semiconductor field-effect transistor (MOSFET) compared to a GaAs field-effect transistor (FET) are (1) the low-resistivity substrate and (2) the high sheet resistance of the polysilicon gates. The substrate resistivity of most deep-submicron silicon processes is 10–15 Ω-cm, which is many orders of magnitude lower than that of GaAs (~10^7–10^9 Ω-cm) [16]. Signals that couple to the low-resistivity silicon substrate incur significant losses, especially at mm-wave frequencies. Furthermore, whereas a GaAs FET can effectively be treated as a three-terminal device, the existence of the bulk terminal and the body-effect complicate matters for MOS modeling and design.

The gate material used for CMOS devices is polysilicon, which has a much higher sheet resistance (~10 Ω/□) than the metal used for the gates of GaAs FETs. A higher gate resistance can reduce the transistor power gain and increase noise figure. Fortunately, simple layout techniques can be used to minimize the detrimental effects of the polysilicon gate.

2.2.2 Back-end features

Fig. 2.1 shows the core back-end stack of the digital CMOS process that was used in this work, consisting of six levels of copper metallization, a low-resistivity silicon substrate, and multiple interlayer dielectrics (oxides and nitrides). Top-layer metal is 0.9 μm thick and the distance from the substrate is 5 μm. Chemical mechanical polishing (CMP) is used to planarize all metals

and dielectrics, providing better repeatability of the conductor and oxide thicknesses compared to GaAs. Due to the use of CMP, though, uniform density is required on all metal levels. Thus, floating dummy fill metal is needed to increase the local density, while large areas of metal (e.g., for ground planes) are forced to have slots.

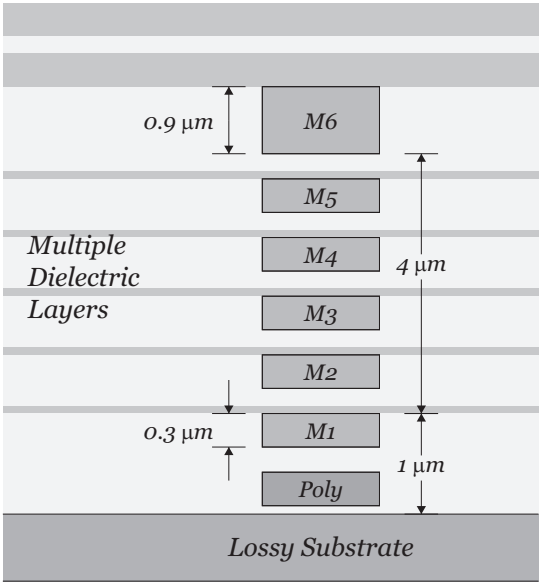


Figure 2.1 Typical CMOS back-end process with six levels of copper metallization, low-resistivity (~10 Ω-cm) silicon substrate, and multiple dielectrics (not to scale).

2.3 CMOS Transistor Design

In this section, the metric used to quantify the high-frequency performance of active devices is first described. Secondly, the effect of parasitics on the high-frequency performance limits of CMOS transistors is presented. Finally, design guidelines for optimal transistor layout are provided along with experimental verification.

2.3.1 Maximum oscillation frequency, f_{max}

For a transistor to be used in the design of all the necessary circuit blocks for a wireless transceiver operating at mm-wave frequencies, it must be able to achieve power gain at the desired frequency of operation. A device is said to be *active* if, under sinusoidal excitation, the net ac power consumed by the network is negative—that is more ac power flows out of the device than into it. At the target operating frequency, the more active the device, the more suitable it will be for circuits at that frequency.

The maximum frequency of oscillation (f_{max}) is generally defined as the frequency that the device transitions from active to passive [17]. Since a circuit composed of only passive components must itself be passive, an immediate consequence is that *all* linear amplifier and oscillator topologies employing this device, including distributed amplifiers, distributed oscillators, and circuits with arbitrary feedback, fundamentally cannot operate beyond f_{max}. Thus, the most relevant figure-of-merit for the high-frequency capabilities of a technology is f_{max}, rather than the transition frequency (f_t), which has been traditionally used as the speed benchmark for digital and baseband circuits.

2.3.2 Mason's unilateral gain

If a transistor is modeled as a linear two-port network, a common figure-of-merit used to characterize the active device is Mason's unilateral gain [18],

$$U = \frac{1}{4}\frac{\left|y_{21} - y_{12}\right|^2}{g_{11}g_{22} - g_{12}g_{21}} \qquad \textbf{(2.1)}$$

where g_{ij} = Re (y_{ij}). One property of the unilateral gain, as proved by Mason, is that U is invariant to lossless reciprocal embeddings. Therefore, the addition of lossless parasitic capacitors or inductors around the device will not change U.

It can be readily shown that, if $g_{11} > 0$ and $g_{22} > 0$, which is true for almost all CMOS transistors, the device is active if

$$U > 1 \tag{2.2}$$

Although (2.2) is often used as the criterion for activity, it should be noted that, strictly speaking, a device can also be active if the denominator

$$g_{11}g_{22} - g_{12}g_{21} < 0 \tag{2.3}$$

The condition in (2.3) is rarely encountered in practice, justifying the widespread use of (2.2) as a test for activity.

In addition to being used as a test for activity, for most practical CMOS transistors, U is a monotonically decreasing function of frequency, and f_{max} is, therefore, the frequency where $U = 1$. Since f_{max} is often larger than the frequency capabilities of the measurement system, it is common practice to use low-frequency measurements of U and report f_{max} as the *extrapolated* frequency where $U = 1$ (often assuming a 20 dB/decade slope). Care must be taken when designing circuits at frequencies approaching the extrapolated f_{max} because U is a complex function of frequency, often dropping at a rate much faster than 20 dB/decade at frequencies near f_{max}. It is imperative, then, that U is measured and modeled as close to the targeted operating frequency of the circuit as possible to minimize errors associated with extrapolation.

One final point should be made about using U to find f_{max}. It is common for other researchers and textbooks to define f_{max} as the frequency that the maximum available gain (MAG)—the power gain achieved with a simultaneous conjugate match at the input and output—becomes unity, instead of U. While it can be proved that the definitions are equivalent (i.e., U = MAG = 1 at f_{max}), there is one major drawback to using MAG instead of U to find f_{max}. While U is defined regardless of stability, the maximum available gain is only defined for unconditionally stable devices. At frequencies where

the device is potentially unstable, the maximum stable gain (MSG)—defined as y_{21}/y_{12}—is used instead of MAG. For deep-submicron CMOS transistors, the frequency that the device becomes unconditionally stable can occur well into the mm-wave regime, making it difficult to directly measure the MAG.

As an example, the unilateral gain, MSG/MAG, and short-circuit current gain for a 100×1 µm/0.13 µm CMOS device biased at 300 µA/µm are shown in Fig. 2.2. For this device, the maximum frequency of oscillation is f_{max} = 135 GHz. This value is much larger than f_t = 80 GHz extrapolated from low-frequency h_{21}, and much lower than the value of 200 GHz extrapolated from low-frequency unilateral gain. The location of the point where the device becomes unconditionally stable (the "knee") occurs at 80 GHz. All f_{max} values reported in this thesis are extrapolated from the circuit models described in Section 2.4 and do *not* assume a 20 dB/decade slope.

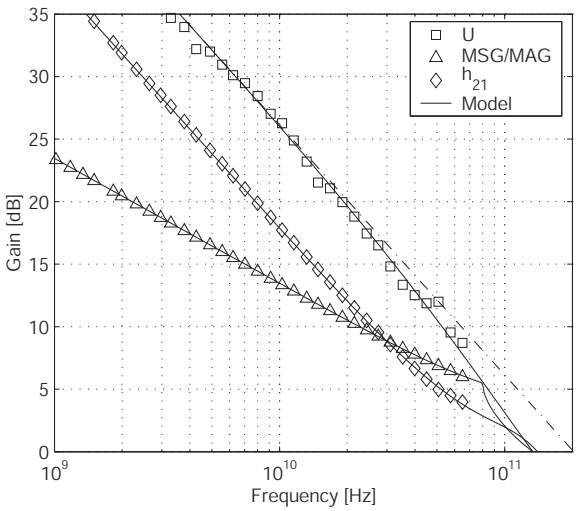


Figure 2.2 Measured (markers) and modeled (solid lines) unilateral gain, maximum stable gain, maximum available gain, and current gain, for a 100×1 µm/0.13 µm NMOS device biased at 30 mA.

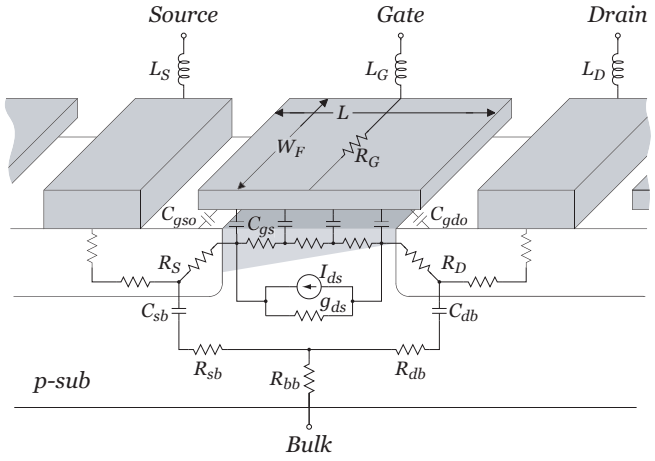


Figure 2.3 Simplified physical model for one finger of an NMOS device.

2.3.3 Layout for optimal f_{max}

Multi-finger transistors Consider a multi-finger transistor composed of N_F identical devices in parallel. If ideal connections are assumed between all of the devices, the port admittance matrix of the composite device is $Y_{TOT} = N_F Y_{Finger}$. Since all of the admittances are scaled by N_F, it is easy to see from (2.1) that

$$U_{TOT} = U_{Finger} \tag{2.4}$$

Therefore, f_{max} of the multi-finger transistor is identical to f_{max} of the individual fingers. Even if non-ideal interconnect is modeled, there is minimal impact on f_{max} because the dominant additional parasitic for transistors with geometries much smaller than a wavelength is additional low-loss parasitic capacitance and inductance. This high-Q parasitic has a negligible effect on U. So, for typical device sizes, it is generally sufficient to only consider the optimal layout for a single finger of a multi-finger transistor.

Optimal finger width The physical layout of a single finger is shown in Fig. 2.3, along with a physical model depicting the dominant high-frequency loss mechanisms. As mentioned, f_{max} is highly-dependent on the resistive losses, the most significant being the gate resistance (R_G), series source/drain resistances (R_S, R_D), non-quasi-static channel resistance (r_{nqs}), and resistive substrate network (R_{sb}, R_{db}, and R_{bb}) [19].

By using narrow finger widths, the effect of the gate resistance can be made small compared to the other parasitic resistors. The polysilicon gate sheet resistance only affects how narrow the fingers must be made. Another benefit of using narrow fingers is that the substrate contacts can be placed very close to the device to minimize the substrate losses.

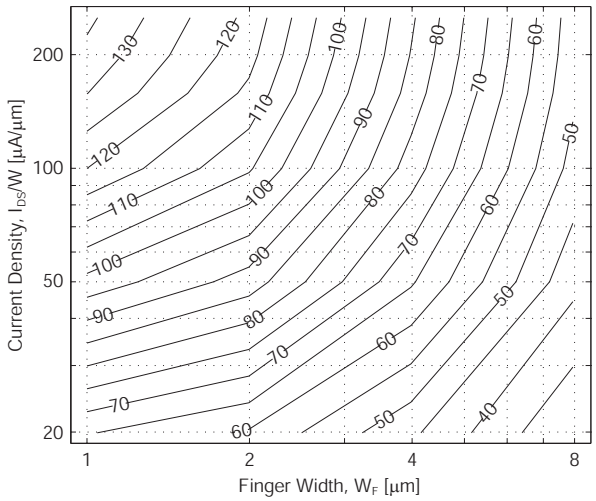


Figure 2.4 Measured maximum frequency of oscillation [GHz] for 130-nm NMOS transistors.

2.3.4 Measured f_{max}

The optimal transistor finger width for our 130-nm digital CMOS process has been determined empirically. The measured f_{max} for NMOS transistors with minimum channel length as a function of finger width and bias current density

is displayed in Fig. 2.4. Nine devices with W_F = 1–8 μm and N_F = 40–100 in common-source configuration have been fabricated. The bulk and source are grounded and the gate contacted on one side for all transistors. For six bias points (I_{DS}/W = 20–300 μA/μm) per device, a transistor model was extracted from the measured data in order to find f_{max} (see Section 2.4). The constant f_{max} contours shown in Fig. 2.4 were linearly interpolated between the measured data points.

For a constant current density, the device f_t remains fixed (e.g., for I_{DS}/W = 100 μA/μm, f_t is 70 GHz). It is clear from Fig. 2.4 that, depending on the finger width, f_{max} can be much larger or smaller than f_t. Thus, the optimal layout for mm-wave applications requires CMOS transistors to be designed using many extremely narrow fingers in parallel (less than 1 μm each). This is in stark contrast to GaAs FETs with metal gates, where relatively few fingers of wide devices ($W_F \approx$ 30–75 μm) are typically used [16]. Furthermore, the device must be biased well into strong inversion (around 100–300 μA/μm) for mm-wave operation. By proper layout and biasing, though, the f_{max} of an NMOS transistor in a standard 130-nm CMOS technology can easily surpass 100 GHz, allowing the possibility for mm-wave amplifier and oscillator circuits.

2.4 Microwave Transistor Modeling

Accurate device models capable of predicting the broadband performance of the transistors are critical for circuits operating near the technology limits. The traditional microwave approach to transistor modeling uses measured *S*-parameter data for circuit design. Although *S*-parameter models are sufficient for many designs and are very accurate, since they implicitly account for all parasitics and any distributed effects, a circuit model provides the ability to

extrapolate to frequencies beyond the measurement capabilities of the test equipment. Additionally, due to de-embedding limitations (Chapter 3), *S*-parameter data for the intrinsic devices become inaccurate at very high frequencies. Finally, an accurate nonlinear large-signal transistor model [20] is required for the design of mixers, oscillators, and power amplifiers.

Noise in RF and mm-wave transceiver circuits plays a critical role in determining the sensitivity of wireless communication systems. With the desire to have highly-integrated CMOS solutions for these applications, there is a strong demand for CMOS transistor models that can accurately predict the device noise up to mm-wave frequencies. At these frequencies, the dominant source of noise is thermal noise, while flicker noise is negligible. In addition to the conventional drain current thermal noise model used for baseband analog circuits, induced gate noise caused by the non-quasi-static (NQS) nature of the channel becomes important. Additionally, the thermal noise generated by the lossy substrate must also be included.

2.4.1 Small-signal modeling

At mm-wave frequencies, series resistive and inductive parasitics become more significant. Consequently, it is critical to properly model these parasitics, in addition to the capacitive effects that are traditionally captured by digital CMOS models [19]. Considering the small margins for modeling errors and the tight coupling between the precise values of the parasitics and the layout details—input and output connections, locations of grounds and substrate contacts, number of fingers, etc.—a transistor modeling approach using the extended circuit model shown in Fig. 2.5 for fixed device layouts is proposed [20]. The core device can be modeled using either a lumped small-signal bias-dependent model for highest accuracy or a standard BSIM3 model card to capture nonlinearities. Notice that all of the capacitors (e.g., C_{gd}, C_{gs}, C_{ds})

account for both the traditionally "intrinsic" channel and overlap capacitances as well as the traditionally "extrinsic" wiring capacitances.

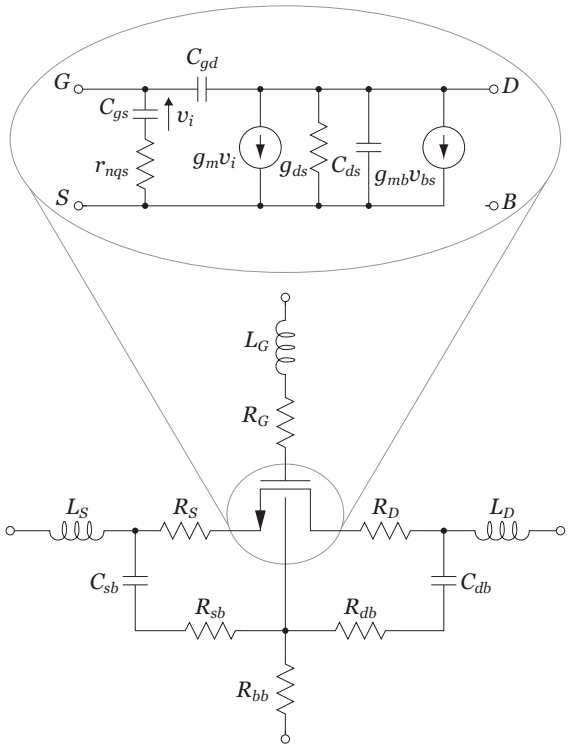


Figure 2.5 Extended transistor model for an NMOS device showing the important parasitic elements.

For each model, the component values and device parameters were extracted from measured data using a hybrid optimization algorithm in Agilent IC-CAP [21]. *S*-parameters for the simulated small-signal model and measured data up to 65 GHz are shown in Fig. 2.6 for a 100×1 μm/0.13 μm NMOS transistor biased at V_{GS} = 0.65 V and V_{DS} = 1.2 V. The excellent broadband accuracy of the simulation compared to the measured data verifies that the topology of the model is correct and complete. Furthermore, it also demonstrates that distributed effects and frequency-dependent losses caused by the skin effect can be adequately accounted for using only lumped extrinsic components with frequency-independent values.

The transistor gains—Mason's unilateral gain, maximum stable gain (*MSG*), maximum available gain (*MAG*), and current gain—for this device are plotted in Fig. 2.2. The accurate modeling of the unilateral gain is particularly important. Unlike the *MSG* and current gain, Mason's unilateral gain is a very strong function of all resistive losses. Therefore, accurately fitting the unilateral gain validates that the important loss mechanisms have been properly modeled. As mentioned earlier, these resistive losses are critical because they ultimately limit the high-frequency capabilities of the transistor.

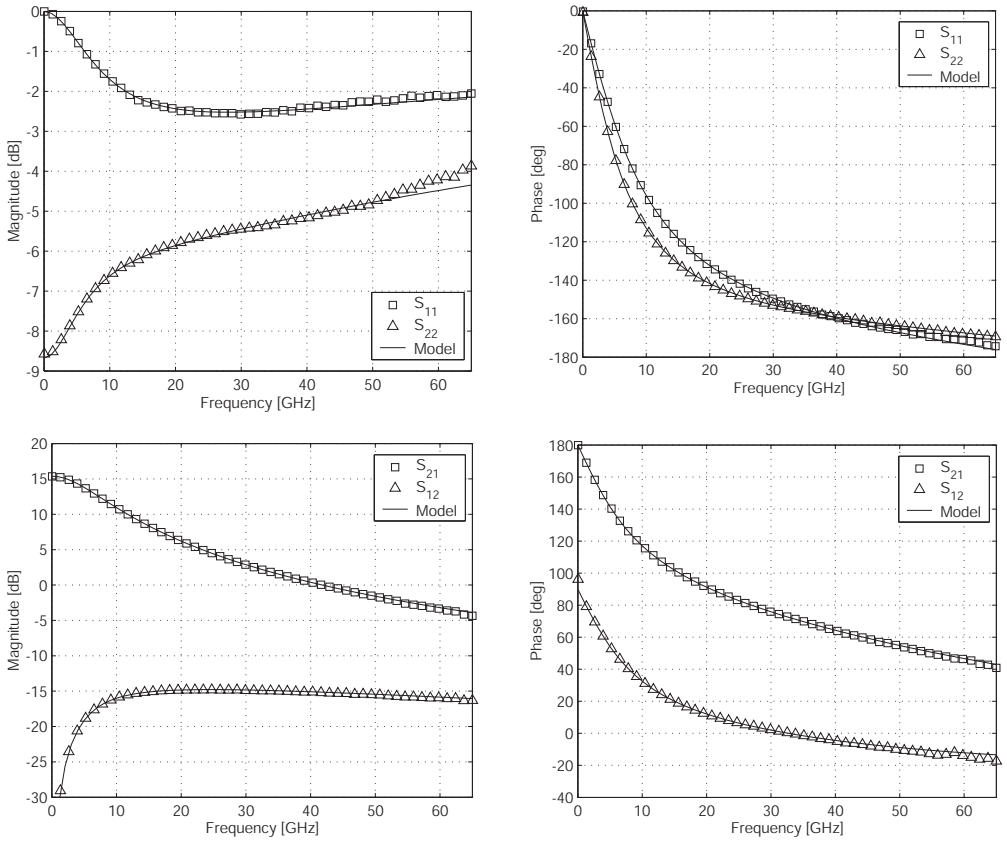


Figure 2.6 Measured (markers) and simulated (solid lines) *S*-parameters for a 100×1 μm/0.13 μm NMOS device biased at 30 mA.

2.4.2 Noise modeling

High-frequency transistor noise The conventional noise model used for baseband analog circuits [22] is not sufficient for short-channel CMOS devices operating at mm-wave frequencies and must be augmented in two significant ways: increased drain current thermal noise and induced gate noise effects [19], [23], [24]. Additionally, at high frequencies, the thermal noise contributed by the same extrinsic resistors that reduce the power gain becomes increasingly important (Fig. 2.3). However, as long as separate elements are used for these physical resistors, their thermal noise contributions will automatically be included in the circuit simulation.

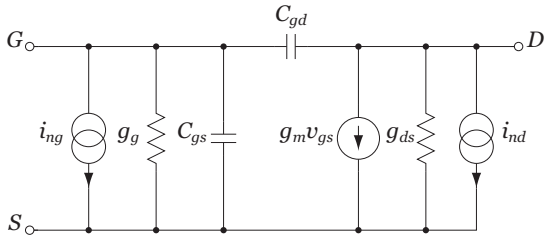


Figure 2.7 Simplified van der Ziel noise model for intrinsic CMOS transistor.

Based on the derivations by van der Ziel, the simplified noise model for the intrinsic CMOS device with drain and gate noise generators is shown in Fig. 2.7. The drain channel noise has a white power spectral density of the form

$$\frac{\overline{i_{nd}^2}}{\Delta f} = 4kT_0\left(\frac{\gamma}{\alpha}\right)g_m \qquad \textbf{(2.5)}$$

where T_0 is the device operating temperature, γ is a bias-dependent "noise enhancement" factor, and α is defined as

$$\alpha \equiv \frac{g_m}{g_{d0}} \le 1 \qquad \textbf{(2.6)}$$

with g_{d0} the drain conductance when there is zero drain-source bias. For long-channel devices biased in saturation, $\gamma = 2/3$. The noise enhancement factor increases for short-channel devices, although there is no clear consensus as to the physical mechanism. One theory attributes the increase to the additional charge build-up near the drain caused by velocity-saturated carriers. Another possible explanation is that the increased noise is due to the existence of hot carriers causing the electron temperature to exceed the lattice temperature. A typical value of the noise enhancement for short-channel CMOS devices operating in saturation is around 1–2 [25].

The channel noise couples to the gate terminal through the gate capacitance, causing a noise current to flow with power spectral density

$$\frac{\overline{i_{ng}^2}}{\Delta f} = 4kT_0\delta g_g \tag{2.7}$$

where g_g is due to the NQS channel resistance, and has a value of

$$g_g = \frac{\alpha\omega^2 C_{gs}^2}{5g_m} \tag{2.8}$$

The factor δ for the noise current is equal to 4/3 for long-channel devices, but can be as large as 15/2 for short-channel devices [26].

Notice that the gate noise current is not white and increases with frequency. However, it can be shown [24] that the shunt noise current can be approximated with an equivalent series noise voltage at the gate, possessing a white power spectral density

$$\frac{\overline{v_g^2}}{\Delta f} = 4kT_0\left(\frac{\alpha\delta}{5g_m}\right) \tag{2.9}$$

when

$$\omega << \frac{5\omega_T}{\alpha} \tag{2.10}$$

Even though the two noise currents arise from the same physical source, the finite response time of the channel (i.e., non-quasi-static (NQS) effect) causes the noise currents to only be partially correlated. The correlation coefficient is defined as

$$c \equiv \frac{\overline{i_{ng} i_{nd}^*}}{\sqrt{\overline{i_{ng}^2} \cdot \overline{i_{nd}^2}}} \tag{2.11}$$

In the long-channel limit, $c \approx j0.395$ [24]. The parameters in (2.5)–(2.11) are generally bias-dependent and frequency-dependent. Attempts to capture these noise effects have led to various approaches for FET noise models, which can broadly be divided into three distinct groups: physics-based compact models, measurement-based empirical models, and equivalent-circuit small-signal models.

Physics-based compact models Intensive research into developing extensions for compact CMOS noise models has received the most attention over the past decade, due to the widespread availability and use of compact models, such as BSIM [32] and Philips MOS Model 11 [25], for digital and baseband analog circuits. BSIM3 does not include RF noise effects such as induced gate noise. However, BSIM4 improves on the RF noise modeling of BSIM3 with a holistic noise model and noise partitioning implementation to account for the increased channel noise and the correlated induced gate noise (Fig. 2.8). The magnitude of the correlation coefficient, but not the complex phase, can be captured using this approach. Philips MOS Model 11 uses a channel segmentation technique, dividing the transistor into many smaller transistors that are individually modeled as quasi-static due to their small size (Fig. 2.9).

In this way, no special considerations are needed for the NQS effects, as they are properly included automatically. Unfortunately, many simulators do not fully support these advanced RF compact models, and more substantially, the foundries typically do not provide these more complicated models. Therefore, noise modeling using compact models was not investigated for this work, although the accuracy of future compact models is promising given the amount of ongoing research in this area.

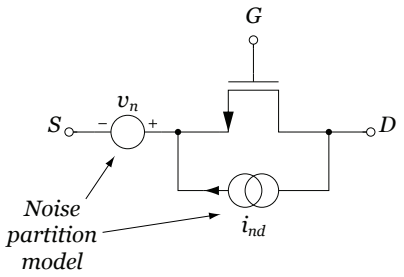


Figure 2.8 BSIM4 noise model.

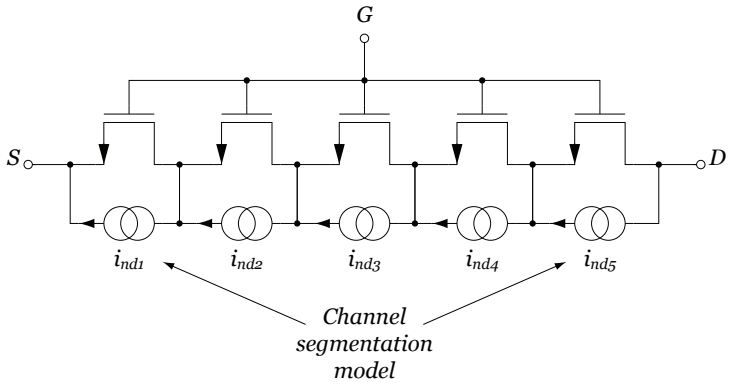


Figure 2.9 Philips MOS Model 11 noise model.

Measurement-based empirical models The general theory behind equivalent circuit models for arbitrary noisy two-ports is described in detail in Section 3.6.1. For a fixed device geometry and bias, the noise characteristics of any two-port device can be completely described with the frequency-dependent

noise parameters (F_{min}, Y_{opt}, R_n). The noise parameters can be determined empirically through measurements and, along with the device *S*-parameters, can be used in the design of low-noise amplifiers [54]. Empirical noise models provide the same benefits and suffer from the same drawbacks as the empirical *S*-parameter small-signal models. The equivalent noise model completely captures the noise behavior of the device without the need to identify the individual noise sources and their correlation to each other. However, large measurement errors at mm-wave frequencies and the inability to extrapolate or predict performance at frequencies other than those measured limit their application.

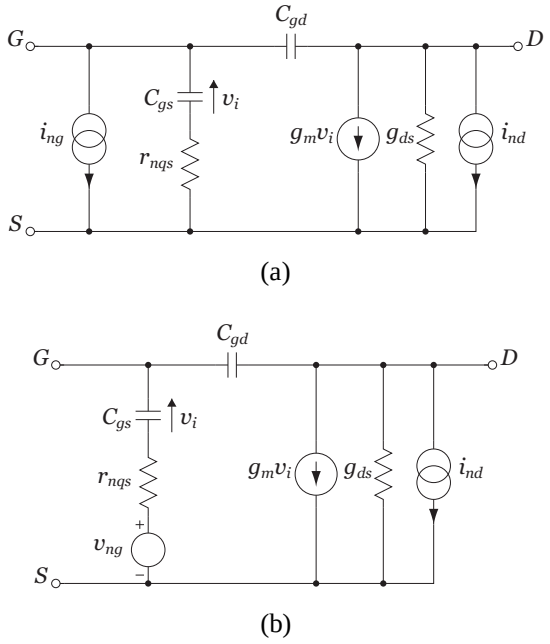


Figure 2.10 Noise model used for (a) PRC and (b) Pospieszalski models.

Equivalent-circuit small-signal models Equivalent-circuit small-signal noise models are a compromise between the physics-based and measurement-based models, and are commonly used in MMIC designs to model the intrinsic device. The two most prevalent models of this type are the *PRC* model [27][28] and Pospieszalski model [29], which are shown in Fig. 2.10.

The two small-signal noise models are very similar. In addition to the minor topology difference, there are a few key distinguishing features. The *PRC* noise model essentially models the noise behavior predicted by van der Ziel using three dimensionless parameters—*P*, *R*, and *C*—to capture the magnitude and correlation of the noise currents. These parameters are defined as

$$\frac{\overline{i_{nd}^2}}{\Delta f} = 4kT_0 g_m \cdot P \tag{2.12}$$

$$\frac{\overline{i_{ng}^2}}{\Delta f} = 4kT_0 \frac{\omega^2 C_{gs}^2}{g_m} \cdot R \tag{2.13}$$

$$\frac{\overline{i_{ng} i_{nd}^*}}{\sqrt{\overline{i_{ng}^2} \cdot \overline{i_{nd}^2}}} = jC \tag{2.14}$$

Comparing these with (2.5)–(2.11), it can be clearly seen that the *PRC* noise model captures the same behavior as the van der Ziel model.

The Pospieszalski noise model is a simplification of the more comprehensive *PRC* model. It models the transistor noise performance using two, rather than three, uncorrelated noise sources specified by the noise temperature of the gate and drain, T_g and T_d, respectively. These are defined as

$$\frac{\overline{v_{ng}^2}}{\Delta f} = 4kr_{nqs} \cdot T_g \tag{2.15}$$

$$\frac{\overline{i_{nd}^2}}{\Delta f} = 4kg_{ds} \cdot T_d \tag{2.16}$$

Equating (2.16) with (2.5), we see that

$$\frac{T_d}{T_0} = \left(\frac{\gamma}{\alpha}\right)\left(\frac{g_m}{g_{ds}}\right) = \left(\frac{\gamma}{\alpha}\right) g_m r_o \tag{2.17}$$

From (2.17), the drain noise temperature can be seen to be much larger than the device operating temperature under most conditions.

Based on empirical data [29][30], it has been found that the gate noise temperature T_g of the Pospieszalski model is often close to the ambient temperature of the device. The physical extrinsic gate resistance R_G (Fig. 2.5) also contributes noise based on the ambient temperature. During model extraction, it is often difficult to partition between the extrinsic gate resistance and the non-quasi-static channel resistance. Therefore, the approximation that $T_g = T_0$ is particularly convenient, since the noise temperature is identical regardless of the physical source. Given this approximation, the simplified Pospieszalski model reduces to being specified using a single parameter T_d.

Experimental verification of noise model Using the on-wafer noise measurement and de-embedding technique described in Section 3.6, the noise parameters were measured from 50–75 GHz for three common-source NMOS transistors and a cascode device. The probe pads were de-embedded such that the noise parameters are for the intrinsic device.

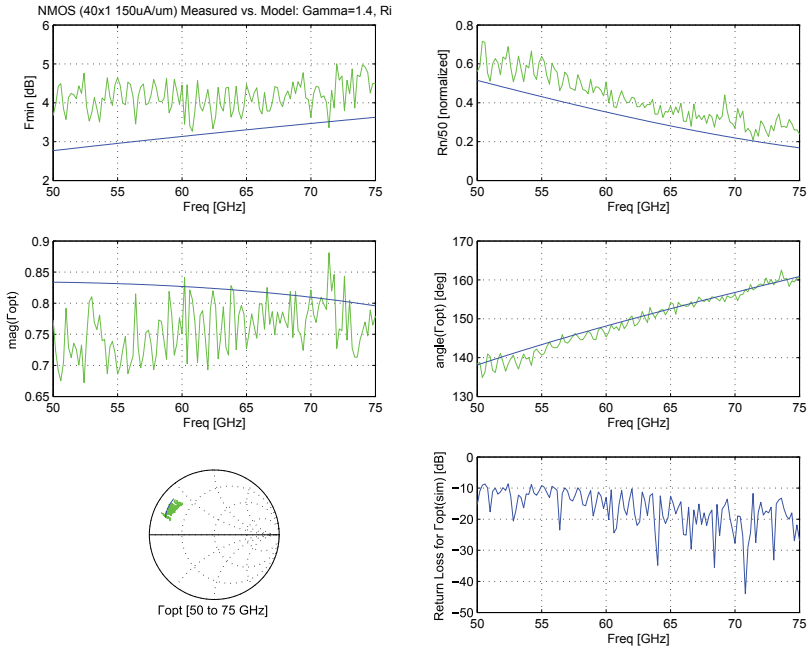
Small-signal models for the devices were extracted based on measured *S*-parameters up to 65 GHz. Based on the simplified Pospieszalski model, aside from the output resistance at the drain, all resistors (including the non-quasi-static channel resistance) introduce thermal noise at a temperature of $T_0 = 290°$ C. The drain noise temperature in (2.17) was determined by defining the parameter

$$\left(\frac{\gamma}{\alpha}\right) = 1.4 \qquad \textbf{(2.18)}$$

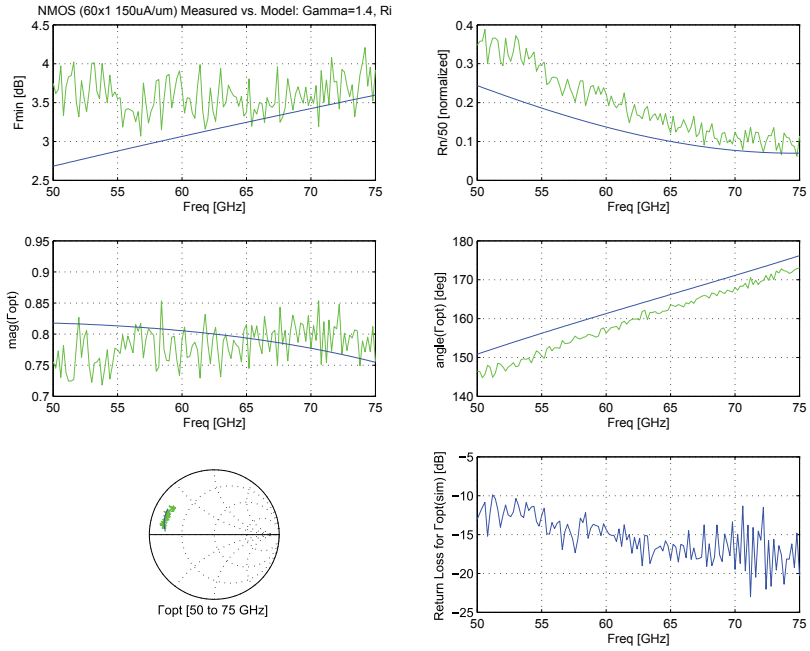
that was chosen to be independent of device geometry. The value for the drain noise enhancement factor in (2.18) was optimized to fit the noise parameters for all devices.

The measured and modeled noise parameters for the devices are shown in Figs. 2.11 and 2.12. Even with the relative simplicity of this noise model, the simulation predicts the noise performance well at mm-wave frequencies. For the single transistors, F_{min} is within 0.5–1 dB across the frequency band, and Γ_{opt} is accurately predicted to better than what corresponds to a 10 dB mismatch. This implies that a low-noise amplifier, designed with a source impedance of Γ_{opt} predicted by the noise model, will still achieve close to the minimum possible noise figure for that device.

It is interesting to note that the minimum noise figure for the cascode device is substantially larger than for a single NMOS transistor, which differs compared to lower frequencies where the cascode noise is often treated as a second-order effect. At frequencies approaching f_{max}, the capacitance at the intermediate node between the two devices causes the gain of the common-source transistor to be lower and also reduces the degeneration for the cascode transistor. Both of these related effects increase the noise contributed by the cascode transistor to the overall noise figure.

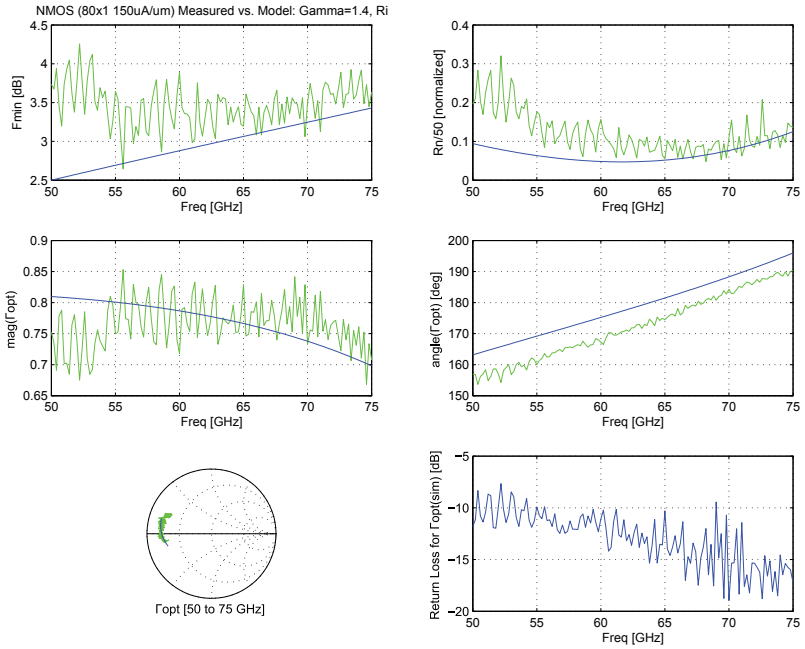


(a)

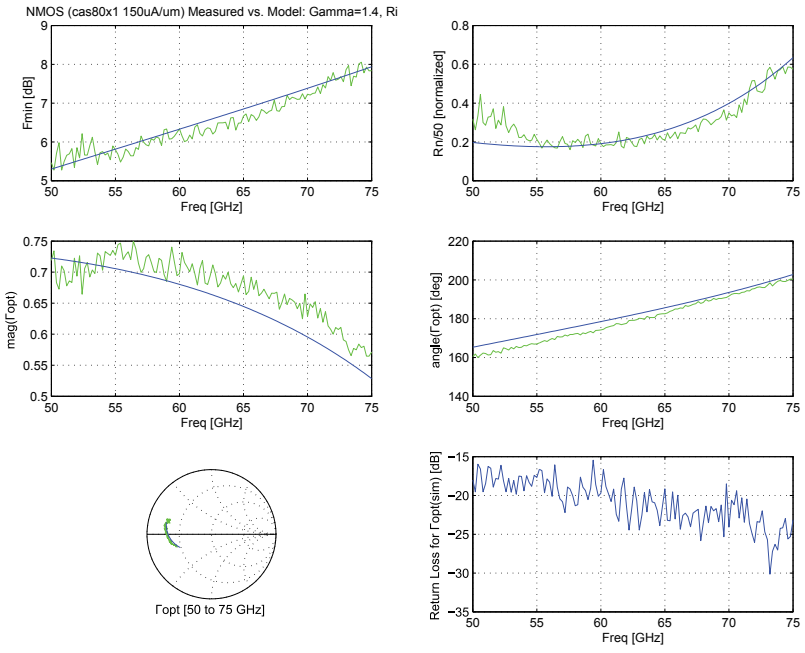


(b)

Figure 2.11 Measurement and simplified Pospieszalski model for minimum noise figure (F_{min}), noise resistance (R_n), and optimal source reflection coefficient (Γ_{opt}) for devices biased at 150 µA/µm. NMOS devices sized at (a) 40×1 µm/0.13 µm (b) 60×1 µm/0.13 µm.



(a)



(b)

Figure 2.12 Measurement and simplified Pospieszalski model for minimum noise figure (F_{min}), noise resistance (R_n), and optimal source reflection coefficient (Γ_{opt}) for devices biased at 150 µA/µm. NMOS devices sized at (a) 80×1 µm/0.13 µm (b) Cascode 40×2 µm/0.13 µm.

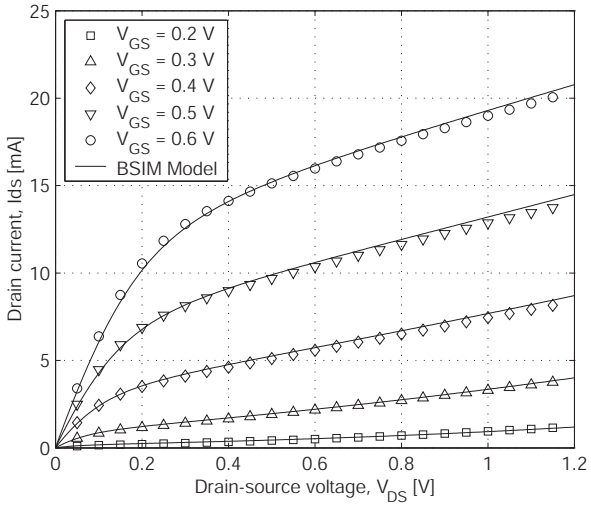
2.4.3 Large-signal modeling

The optimization of nonlinear circuits such as mixers, power amplifiers, oscillators, and frequency multipliers, requires precise knowledge of the nonlinear characteristics of the active devices over a wide range of operation. Large-signal device models have evolved into two basic categories: table-based [31] and physical [32][33]. At mm-wave frequencies, GaAs FET models commonly employ table-based models derived from bias-dependent linear measurements. The large-signal accuracy of table-based models is limited by the existence of discontinuities in the model elements and nonlinearities in their interpolation due to imperfect measurement data [34].

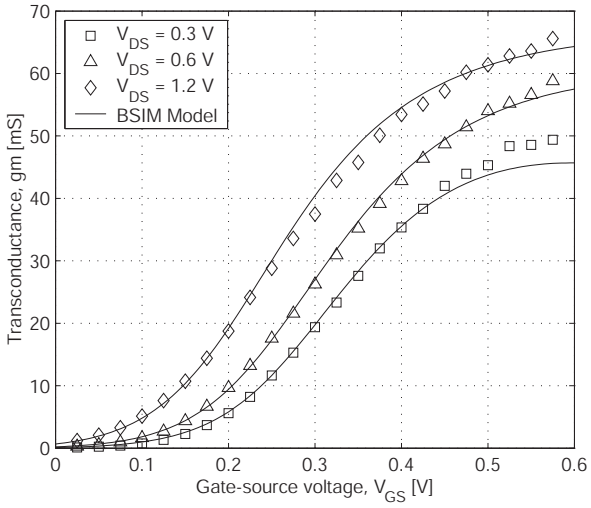
We have developed a large-signal modeling methodology for fixed geometry devices [20] that is based upon a standard intrinsic BSIM3v3 model augmented with bias-dependent junction capacitors and lumped elements to model the dominant mm-wave effects. The proposed methodology is founded on the quasi-static assumption that the mm-wave large-signal performance of a transistor is primarily governed by its dc nonlinearities, while the nonlinear capacitors only contribute a small amount to the high-frequency distortion. The dynamic performance is captured with the addition of extrinsic parasitics to capture loss and inductive effects, as was discussed in Section 2.4.1.

The core BSIM3v3 parameters were first extracted to match the measured dc I-V curves of the fabricated common-source NMOS transistors. As shown in Fig. 2.13, a good agreement between measured and modeled dc curves can be achieved for this device (80×1 μm/0.13 μm NMOS). Measured *S*-parameters are then used to extract external parasitic component values to obtain a bias-dependent small-signal mm-wave frequency fit up to 65 GHz. Layout-dependent interconnect capacitances are also added around the intrinsic device, and junction diodes account for the voltage-dependence of C_{db} and C_{sb}. A comparison of measured and simulated data for the de-embedded 80×1

μm/0.13 μm NMOS transistor over a typical bias sweep of V_{GS} is shown in Fig. 2.14, again demonstrating good broadband accuracy.



(a)



(b)

Figure 2.13 Measured and modeled dc nonlinearities. (a) I_{DS} vs. V_{DS}. (b) g_m vs. V_{GS}.

The proposed BSIM3v3 model was implemented in Agilent EEsof ADS, and the harmonic balance simulator was used to predict the large-signal behavior. Harmonic distortion tests, using the measurement setup shown in

Section A4.2, have been performed over numerous operating regions and frequencies with good results, and described in detail in [20]. One particularly important test is when the device is biased near the threshold voltage and operating in Class B mode. Many mm-wave circuits, such as mixers and frequency multipliers, function by exploiting the strong nonlinearity of transistors biased near the threshold voltage, effectively using the device as a rectifier. To determine the accuracy of the model for these applications, the transistor was biased at constant V_{GS} = 0.2 V, and the input power was swept from P_{in} = –3 dBm to +3 dBm at a fundamental of 30 GHz. The results shown in Fig. 2.15 show that the extended BSIM3v3 model provides good accuracy for Class B operation. Thus, the extended BSIM3v3 model is capable of predicting bias-dependent small-signal performance and large-signal distortion performance of the devices up to 65 GHz, and can be used to design nonlinear mm-wave transceiver circuit blocks.

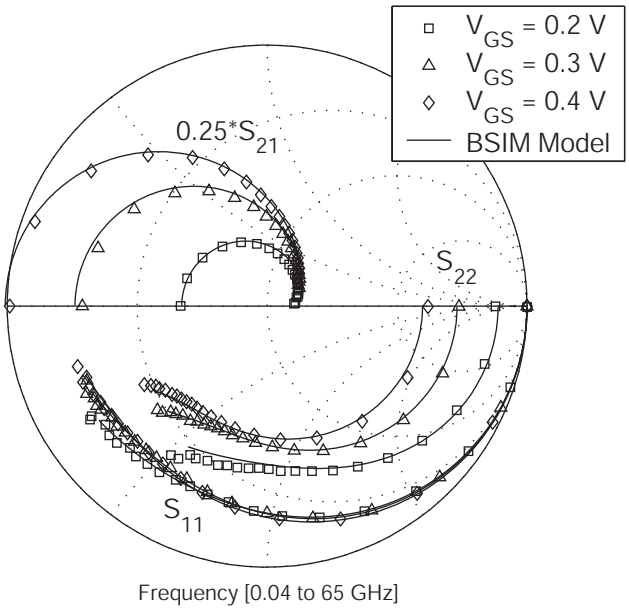


Figure 2.14 Measured and modeled *S*-parameters for V_{DS} = 1.2 V.

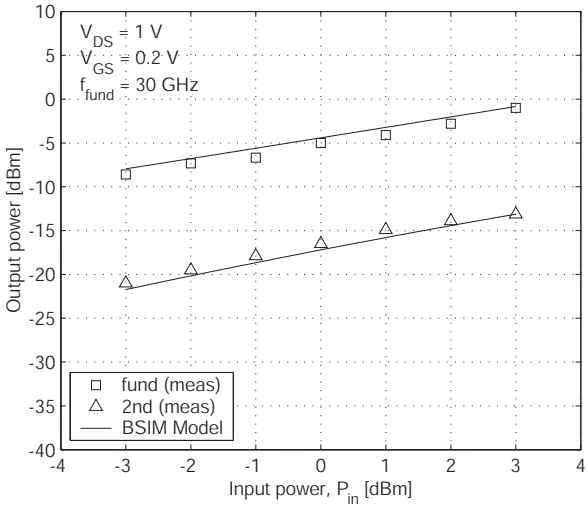


Figure 2.15 Class B power sweep curves.

2.5 Transmission Lines

Transmission lines (T-lines) are important structures for mm-wave design. At these frequencies, the reactive elements needed for matching networks and resonators become increasingly small, requiring inductance values on the order of 50–250 pH. Given the dominant quasi-transverse electromagnetic (quasi-TEM) mode of propagation, T-lines are inherently scalable in length and are capable of realizing precise values of small reactances. Additionally, interconnect wiring can be modeled directly when implemented using T-lines. Another benefit of using T-lines is that the well-defined ground return path significantly reduces magnetic and electric field coupling to adjacent structures.

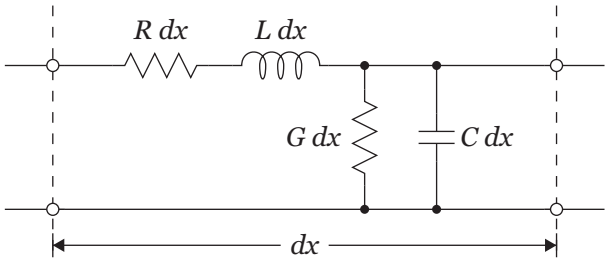


Figure 2.16 Distributed *RLGC* lossy transmission line model.

2.5.1 Characterizing low-loss transmission lines

Four real (or two complex) parameters are needed to completely capture the properties of any quasi-TEM transmission line at a given frequency, ω_0. The T-line can be characterized by its equivalent frequency-dependent *RLGC* distributed circuit model (Fig. 2.16), which can be related to the characteristic impedance (Z_0) and complex propagation constant (γ) by [35]

$$Z_0 = \sqrt{\frac{R + j\omega_0 L}{G + j\omega_0 C}} \tag{2.19}$$

$$\begin{aligned} \gamma &= \sqrt{(R + j\omega_0 L)(G + j\omega_0 C)} \\ &= \alpha + j\beta \end{aligned} \tag{2.20}$$

For low-loss lines, the attenuation and phase constants, α and β, respectively, can be approximated as

$$\alpha \approx \frac{R}{2Z_0} + \frac{GZ_0}{2} \tag{2.21}$$

$$\beta = \frac{2\pi}{\lambda_g} \approx \omega_0\sqrt{LC} \tag{2.22}$$

where λ_g is the wavelength of a quasi-TEM wave on the line. The

characteristic impedance and wavelength are particularly important for design, and therefore, (2.19) and (2.20) are used more often than the *RLGC* model directly, since Z_0 and λ_g are readily observed.

There are two deficiencies with the standard representation of (2.19) and (2.20). First, as seen in (2.21), the two different loss mechanisms are combined into one parameter, making it difficult to discern the relative importance of *R* and *G*. As opposed to transmission lines implemented on a semi-insulating substrate, where *G* is nearly zero, T-lines on low-resistivity silicon have a non-negligible *G* due to the substrate coupling. Secondly, Z_0 is often approximated as a real number, and the expression for $\mathrm{Im}(Z_0)$, which will be shown to be important, is not obvious from this representation.

To address these two issues, the following four *real* parameters are proposed to characterize the line:

$$Z \equiv \sqrt{L/C} \tag{2.23}$$

$$\lambda \equiv \frac{2\pi}{\omega_0 \sqrt{LC}} \tag{2.24}$$

$$Q_L \equiv \omega_0 L/R \tag{2.25}$$

$$Q_C \equiv \omega_0 C/G \tag{2.26}$$

Notice that the two loss mechanisms are completely decoupled in (2.25) and (2.26). If the quantities in (2.23)–(2.26) are related to those of (2.19)–(2.22), the first-order Taylor series expansions are

$$Z_0 \approx Z\left(1+\frac{j}{2}\left(\frac{1}{Q_C}-\frac{1}{Q_L}\right)\right) \tag{2.27}$$

$$\lambda_g \approx \lambda \tag{2.28}$$

From (2.27), the expression for $\mathrm{Im}(Z_0)$ is now apparent. For lossless lines,

$\mathrm{Im}(Z_0) \rightarrow 0$ as expected, whereas for low-loss lines, the sign of $\mathrm{Im}(Z_0)$ reveals which loss mechanism (Q_L or Q_C) is dominant. If Z_0 is assumed to be real, this implicitly requires that $Q_L = Q_C$, which is not generally true. The fact that $\mathrm{Im}(Z_0)$ is non-zero and frequency-dependent causes difficulties for some de-embedding methods, and the implications will be discussed further in Section 3.4.2.

2.5.2 Inductive quality factor

Transmission lines are often used to resonate with the intrinsic capacitance of the transistors (e.g., when used in matching networks). In this case, the line stores mostly magnetic energy, and as described in Section A2.1, it is most appropriate to minimize the power lost for a given amount of net reactive energy stored in the line, as opposed to the total stored energy [36].

$$Q_{net} \equiv 2\omega_0 \frac{\text{net energy stored}}{\text{avg. power loss}} = \frac{2\omega_0 (W_m - W_e)}{P_L} \tag{2.29}$$

where ω_0 is the resonance frequency, W_m and W_e are the average magnetic and electric energy stored, and P_L is the average power dissipated in the line. As derived in Section A2.1, this can be expressed as

$$\frac{1}{Q_{net}} = \frac{1}{\eta_L Q_L} + \frac{1}{\eta_C Q_C} \tag{2.30}$$

where

$$\eta_L = 1 - \frac{W_e}{W_m} \tag{2.31}$$

$$\eta_C = \frac{W_m}{W_e} - 1 \tag{2.32}$$

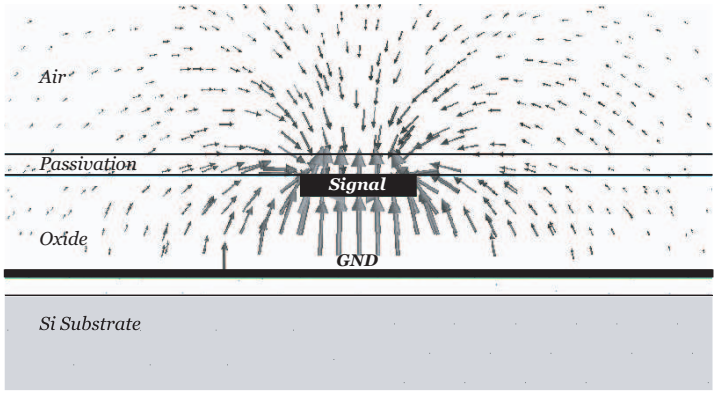
If the line is inductive (i.e., $W_m >> W_e$), then $\eta_C >> \eta_L$ and $Q_{net} \approx \eta_L Q_L$. Thus,

when used to store mostly magnetic energy, the *inductive quality factor*, Q_L, is a better metric than the resonator quality factor as it more accurately corresponds to the amount of loss due to the line.

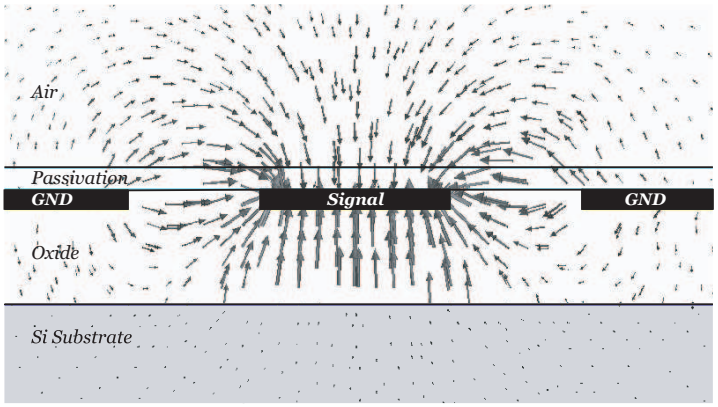
For example, consider a shorted transmission line with $l < 0.1\lambda$. In this case, it can be shown (Section A2.2) that $\eta_C > 7.2\eta_L$, and the relative importance of the shunt losses is reduced compared to the series losses. This is particularly important for integrated T-lines on silicon, where the low-resistivity substrate causes Q_C to be non-negligible. A similar qualitative discussion and conclusion for inductive lines has also been presented in [37].

2.5.3 Microstrip vs. coplanar waveguide

Microstrip lines on silicon are typically implemented using the top-layer metal as the signal line, and the bottom-layer metal for the ground plane. A major advantage of microstrip lines is the shielding it provides from the lossy silicon substrate. Fig. 2.17a illustrates the effectiveness of the metal shield, with essentially no electric field penetration into the substrate. The shunt loss, G, is therefore due only to the loss tangent of the oxide and some radiation loss, yielding a capacitive quality factor, Q_C, of around 30 at mm-wave frequencies (Fig. 2.18b). In addition to reducing the shunt loss, the isolation provided by the ground plane also makes microstrip lines less sensitive to the processing details of the substrate, which in reality is not a simple uniformly-doped semiconductor.



(a)



(b)

Figure 2.17 Electric field distributions from 3-D EM simulations of (a) microstrip and (b) coplanar waveguide transmission lines.

The biggest drawback to microstrip lines on standard CMOS is the close proximity of the ground plane to the signal line (~4 μm), yielding very small distributed inductance, L. This significantly degrades the inductive quality factor, Q_L (Fig. 2.18a). Another disadvantage of microstrip lines is their low characteristic impedance, which signifies large capacitive loading. For example, the characteristic impedance of a 6-μm-wide microstrip line was measured to be only 49 Ω. The Z_0 can be increased by reducing the conductor width, but very high-impedance microstrip lines are impractical since electromigration design rules limit the minimum metal width.

Another option for on-chip transmission lines is the use of coplanar waveguides (CPWs) [38], [39], which are implemented with one signal line surrounded by two adjacent grounds (Fig. 2.17b). The signal width, W, can be used to minimize conductor loss, while the signal-to-ground spacing, S, controls the Z_0 and the tradeoff between Q_L and Q_C. As an example, a CPW with W = 10 μm and S = 7 μm has a Z_0 of 59 Ω and a Q_L measured to be about double that of the microstrip (Fig. 2.18a). By varying the signal-to-ground spacing, it is possible to design CPW lines to have either large Q_L and high-impedance (S = 7 μm) or large Q_C and low-impedance (S = 2 μm) (Fig. 2.19). On the other hand microstrip lines have, to first-order, constant Q_L and Q_C independent of geometry. CPW parameters should also have less variations than microstrip, since the dimensions of a CPW are determined by lithography and not oxide thicknesses.

Although the shielding is not as good, Fig. 2.17b illustrates that most of the electric fields still terminate on the coplanar grounds instead of the lossy substrate. Furthermore, the coupling to the substrate manifests itself as a reduction in Q_C (Fig. 2.18b), which was shown to be relatively insignificant for many circuit applications. Another important issue when designing with CPWs is the unwanted odd CPW mode, which arises because CPW lines inherently have three conductors. To suppress this parasitic propagation mode, the two grounds should be forced to the same potential [38]. In MMICs, this requires the availability of air bridge technology, which is costly and not supported by all foundries. Underpasses using a lower metal level in a modern CMOS process can be used to suppress this mode.

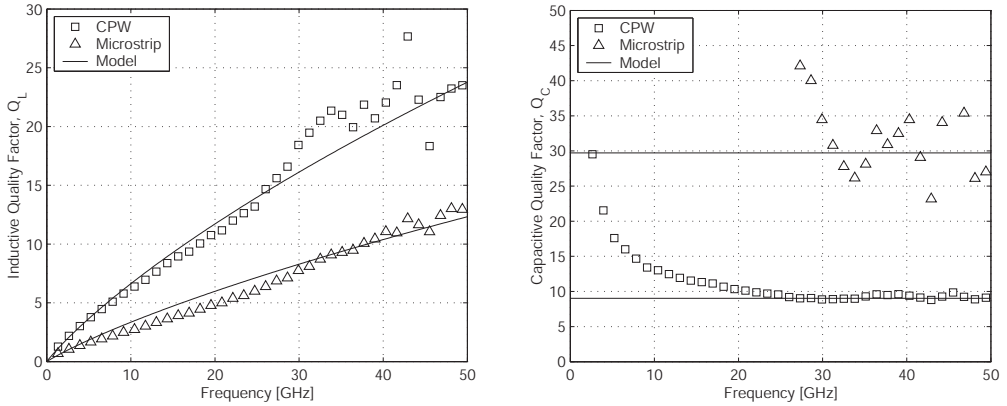


Figure 2.18 Measured (markers) and modeled (solid lines) quality factors for a coplanar waveguide and a microstrip line.

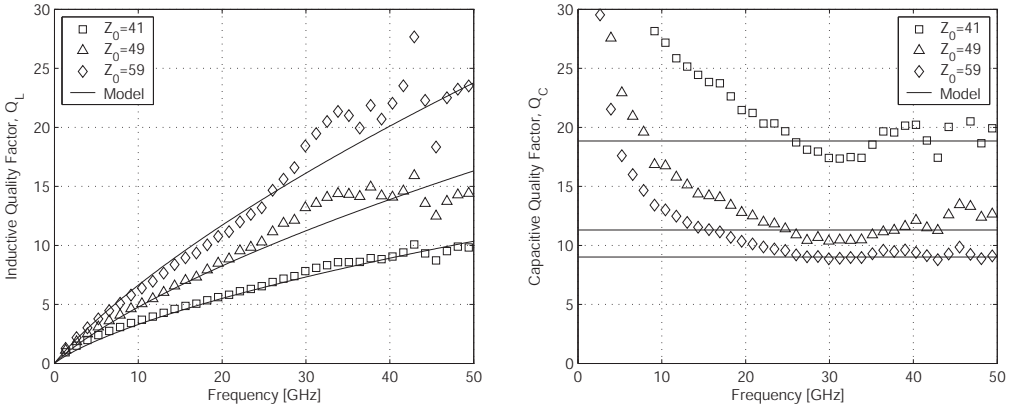


Figure 2.19 Measured (markers) and modeled (solid lines) quality factors for coplanar waveguides with varying geometries.

The ultimate choice of transmission line structure depends upon the application and the specifics of the back-end process. For mm-wave transceivers in this CMOS technology, the considerably higher Q_L for CPW lines outweighs the benefits of shielding afforded by using microstrip lines. Therefore, CPW transmission lines were used in all of our circuit designs.

2.6 Transmission Line Modeling

Accurate transmission line models are imperative, as the T-lines are used extensively in mm-wave circuit designs. Owing to the quasi-TEM nature of propagation, scalable (in length) transmission lines are easier to model than lumped passives. Furthermore, because of the field confinement, the lines can be modeled individually without concern that adjacent structures will affect the performance characteristics of the line when used in more complex circuits.

Equation-based models derived for microwave transmission lines are not applicable for lines integrated on silicon substrates. The models typically assume thin conductors, ground planes which are far from the signal lines, and high-quality dielectrics. The lossy silicon substrate is very close to the signal lines (~5 μm) and the metal thickness is on the order of the dimensions of the conductor width and signal-to-ground spacing. In this section, two modeling approaches for transmission lines are investigated and their trade-offs presented [40].

2.6.1 Physical electromagnetic modeling

Electromagnetic (EM) simulations of the passives based on the physical layout were performed using Ansoft HFSS (Fig. 2.17). Several simplifications were used to improve simulation time with a small degradation in accuracy. The silicon substrate is modeled as uniformly doped, and the multilayer dielectrics are modeled with two layers: the oxide and passivation with effective dielectric constants of 4 and 6.3, respectively. Additionally, the dielectric loss tangent, which degrades with frequency, is assumed to be constant. Although 3-D EM simulations enable the characterization of arbitrary passive devices, they suffer from long simulation times (many hours) and the inability to include several poorly characterized effects (substrate doping profile, conductor surface

roughness, etc.). Physical EM simulations should be used to compare and optimize transmission line cross-sectional geometries, and can serve as a good verification tool for more complex passive networks, but are too slow to be used iteratively during the design phase.

2.6.2 Design-oriented electrical modeling

To achieve a high level of accuracy for the transmission line models when data from fabricated T-lines are available, a design-oriented modeling methodology, similar to [41], has been chosen for this work. The modeling approach is based on the measured transmission line data and the models are optimized to fit the data most accurately at mm-wave frequencies. Scalable (in length) electrical models, which capture the high-level behavior of the lines, have been used and are supported in most simulators such as SpectreRF, ADS, and Eldo. The model parameters are easy to obtain from measured data or physical EM simulations since only a relatively small number of parameters are required to model the broadband performance of each transmission line: characteristic impedance, effective dielectric constant, attenuation constant, and loss tangent. A first-order frequency-dependent loss model is used. The model assumes that the conductor loss is only caused by the skin effect losses, and the shunt loss is due to a constant loss tangent.

Using simple electrical models has many advantages. The simulation time is very fast, and the models can be easily integrated into circuit simulators and optimizers. The T-line models assume no coupling to adjacent structures. This assumption is justified as the well-defined ground return path helps confine the EM fields. In addition, the close proximity of the adjacent grounds to the signal line helps to minimize second-order effects due to bends, junctions, end-effects, radiation, and other discontinuities. These second-order effects were not modeled, and as verified in Chapter 4, this simplification

results in a negligible reduction in accuracy.

A CMOS test chip was fabricated which included 1-mm long CPW and microstrip transmission lines of different cross-sectional dimensions. For each geometry, a different transmission line model was extracted. Figs. 2.18 and 2.19 demonstrate good broadband modeling of the loss using the electrical models for both CPW and microstrip transmission lines.

The somewhat noisy measurement data for the transmission lines can be attributed to two factors. Although the overall attenuation constant of the T-line can be accurately extracted from measurements (Section 3.5.4), decomposing this loss into individual loss components causes the measured high-Q data (e.g., Q_C for microstrip) to exhibit more measurement uncertainty than the low-Q data. Above 45–50 GHz, de-embedding errors result in invalid transmission line data (Section 3.5.4), and therefore, the data is plotted only to 50 GHz.

2.7 Conclusion

Designing circuits for mm-wave operation requires optimized design, layout, and modeling of all the active and passive building blocks. The key parasitics that limit CMOS transistor performance at mm-wave frequencies are discussed and techniques to optimize the layout for best performance are presented. Extended transistor models that capture the gain, noise, and linearity performance are described along with experimental validation up to 65 GHz. Transmission lines are widely used in mm-wave circuits for matching networks and interconnect. Different transmission lines and their models are compared with silicon results. With simulation models that accurately match measured performance up to mm-wave frequencies, it is now possible to design circuits based on these components that can operate near the frequency limits set by the technology.

Appendix

A2.1 Q-FACTOR DEFINITIONS FOR TRANSMISSION LINES

Analogous to the case for lumped inductors [36], there are many different definitions for Q of a transmission line, where the applicability depends upon the intended function of the transmission line in the circuit.

A2.1.1 Resonator quality factor

The most commonly used definition is the Q of the line when used as a resonator (Q_{res}),

$$Q_{res} \equiv \omega_0 \frac{\text{avg. energy stored}}{\text{avg. power loss}} = \frac{\omega_0 (W_m + W_e)}{P_L} \tag{2.33}$$

where ω_0 is the resonance frequency, W_m and W_e are the average magnetic and electric energy stored, and P_L is the average power dissipated in the line. Q_{res} can be related to the quantities introduced in (2.21) and (2.22) by [42]

$$Q_{res} \approx \frac{\beta}{2\alpha} = \frac{\pi}{\alpha \lambda_g} = \frac{\pi \sqrt{\varepsilon_{eff}}}{\alpha \lambda_0} \tag{2.34}$$

where λ_0 is the free-space wavelength and ε_{eff} is the effective dielectric constant. Since ε_{eff} is determined mostly by the dielectric properties, and not the transmission line structure or dimensions, maximizing Q_{res} is roughly equivalent to minimizing the attenuation constant α. Using (2.34) along with (2.21) and (2.22), Q_{res} can be related very simply to the parameters in (2.25) and (2.26) by

$$\frac{1}{Q_{res}} \approx \frac{1}{Q_L} + \frac{1}{Q_C} \tag{2.35}$$

A2.1.2 Net quality factor

If the line is designed to behave as an equivalent inductor, storing mostly magnetic energy, it is more appropriate to consider the power dissipated for a given amount of *net* reactive energy stored in the line, as opposed to the *total* stored energy [36].

$$Q_{net} \equiv 2\omega_0 \frac{\text{net energy stored}}{\text{avg. power loss}} = \frac{2\omega_0(W_m - W_e)}{P_R + P_G} \tag{2.36}$$

where ω_0 is the resonance frequency, W_m and W_e are the average magnetic and electric energy stored, and P_R and P_G are the average power dissipated in the resistance and conductance, respectively. If Q_L and Q_C are expressed as

$$Q_L = 2\omega_0 W_m / P_R \tag{2.37}$$

$$Q_C = 2\omega_0 W_e / P_G \tag{2.38}$$

it is straightforward to show that

$$\frac{1}{Q_{net}} = \frac{1}{\eta_L Q_L} + \frac{1}{\eta_C Q_C} \tag{2.39}$$

where

$$\eta_L = 1 - \frac{W_e}{W_m} \tag{2.40}$$

$$\eta_C = \frac{W_m}{W_e} - 1 \tag{2.41}$$

A2.2 Q-FACTOR OF A SHORTED TRANSMISSION LINE

For a shorted transmission line (Fig. 2.20), Q_{net} is simply the single-ended port Q, which is

$$Q_{net} = -\frac{\mathrm{Im}(y_{11})}{\mathrm{Re}(y_{11})} \tag{2.42}$$

By finding a relationship between Q_{net} of a shorted line and the parameters of (2.23)–(2.26), it will be shown that Q_{net} is almost solely determined by Q_L, and in many instances Q_C can be ignored. This is unlike (2.35), where Q_{res} is equally dependent upon both Q_L and Q_C.

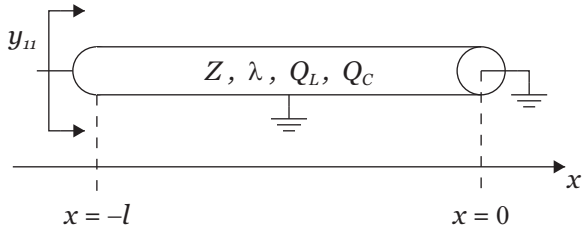


Figure 2.20 Shorted transmission line.

To derive the expression for Q_{net}, the equations for W_m and W_e must first be analytically derived. Then the results of (2.39)–(2.41) can be used to determine the scaling factors η_L and η_C. If the location of the short is at $x = 0$ (Fig. 2.20), then for an ideal lossless line, the voltage and current phasors on the line, as a function of position x, are [42]

$$V = -2jV_+ \sin \beta x \tag{2.43}$$

$$I = 2\frac{V_+}{Z_0} \cos \beta x \tag{2.44}$$

where V_+ is the voltage in the positively traveling wave. The time origin is selected such that V_+ is real.

For a low-loss transmission line of length l, if we assume that the current and voltage are not significantly affected by the losses on the line, then the average stored magnetic and electric energy are

$$W_m = \int_{-l}^{0} \frac{L|I|^2}{4} dx \approx \tfrac{1}{2} \frac{LV_+^{\,2} l}{Z^2} \left(1 + \text{sinc}\left(4l/\lambda\right)\right) \quad \textbf{(2.45)}$$

$$W_e = \int_{-l}^{0} \frac{C|V|^2}{4} dx \approx \tfrac{1}{2} CV_+^{\,2} l \left(1 - \text{sinc}\left(4l/\lambda\right)\right) \quad \textbf{(2.46)}$$

where

$$\text{sinc}\left(t\right) \equiv \frac{\sin\left(\pi t\right)}{\pi t} \quad \textbf{(2.47)}$$

Substituting (2.45) and (2.46) into (2.40) and (2.41), we find

$$\frac{1}{Q_{net}} = \frac{1}{\eta_L Q_L} + \frac{1}{\eta_C Q_C} \quad \textbf{(2.48)}$$

where

$$\frac{1}{\eta_L} = \frac{1}{2\text{sinc}\left(4l/\lambda\right)} + \frac{1}{2} \quad \textbf{(2.49)}$$

$$\frac{1}{\eta_C} = \frac{1}{2\text{sinc}\left(4l/\lambda\right)} - \frac{1}{2} \quad \textbf{(2.50)}$$

For transmission lines that are much shorter than a wavelength, the shorted stub looks inductive, $\eta_C >> \eta_L$, and Q_C can often be neglected. For example, if $l < 0.1\lambda$ then $\eta_C > 7.2\eta_L$.

3

CMOS On-Wafer De-embedding

3.1 Introduction

As described in the previous chapter, accurate modeling of both active and passive devices is the cornerstone of predictable circuit design. The extraction of a good device model is predicated on the ability to perform accurate device measurements against which the model is compared. Poor correspondence between circuit measurements and simulation models is just as easily due to errors in the mm-wave measurements used to generate the models as to the models themselves.

Accurate well-known calibration techniques for small-signal measurements are available that set the reference plane at the wafer probe tip. Unfortunately, it is not possible to directly probe and measure the intrinsic devices. Instead the device-under-test (DUT) is embedded within a test fixture consisting of probe pads and interconnects to the DUT (Fig. 3.1). As the frequency of measurement moves from RF to microwave to mm-wave, fixture

de-embedding becomes increasingly important and challenging, particularly since the pad parasitics are usually much larger than the small devices used for modeling. Therefore, understanding and designing the probe pads and interconnects are keys to accurate device modeling.

This chapter describes the challenges involved in measuring and de-embedding fixtures used to characterize the *S*-parameters and noise parameters of devices fabricated on lossy silicon substrates. Design considerations for CMOS pads that enable accurate de-embedding will be discussed. Several on-wafer calibration and de-embedding techniques and the problems that limit their suitability for characterizing the *S*-parameters of CMOS devices will be described. A verification technique based on TRL calibration is developed and proposed to help identify the validity range of the de-embedding approaches. Finally, the methodology to de-embed the pads from the intrinsic device for device noise characterization will be described.

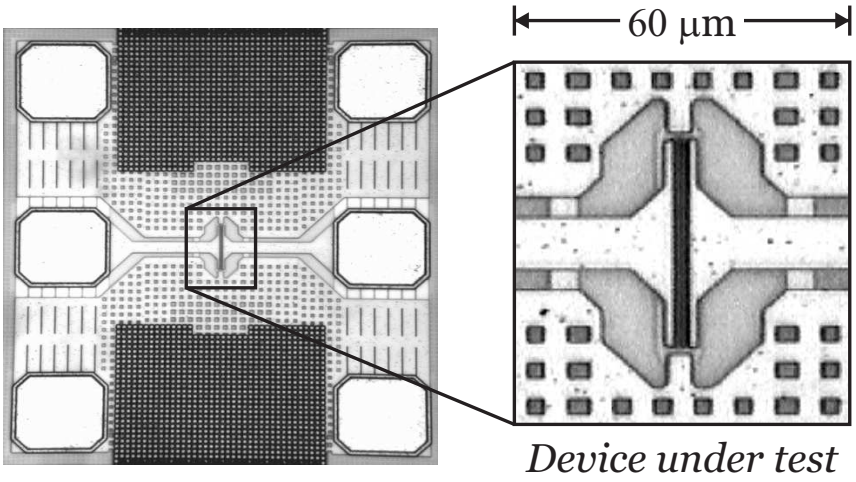


Figure 3.1 Typical layout for a test fixture and DUT.

3.2 Accurate *S*-parameter Measurements on Silicon

Accurate de-embedding is founded on the basic ability to make repeatable and precise on-wafer mm-wave measurements [43]. Commercial impedance standard substrates (ISS) fabricated on alumina can be used to define the measurement reference plane at the probe tips using standard vector network analyzer (VNA) calibration techniques, such as short-open-load-through (SOLT), line-reflect-reflect-match (LRRM), or through-reflect-line (TRL). Ideally, the electrical performance of the probe pads on the ISS should be identical to the on-wafer probe pads, otherwise the interaction between the probe pads and wafer probe will limit the accuracy of the calibration standard. For fixtures fabricated in CMOS, the pads will inevitably be different than those on the ISS; however, this source of error is usually small compared to errors resulting from inconsistent probe placement and overdrive, which change the pad inductance and contact resistance.

Aluminum pads fabricated on CMOS will oxidize, resulting in large contact resistance. In the past, the probe tips were made of tungsten (W) that, due to its firmness, could break through the oxide of the aluminum pads. However, W itself oxidizes and the contact resistance increases over time. Fortunately, nickel-alloy tips have been developed that provide very low and stable contact resistance when probing aluminum pads [44]. This is critical since the contact resistance must be repeatable for accurate de-embedding.

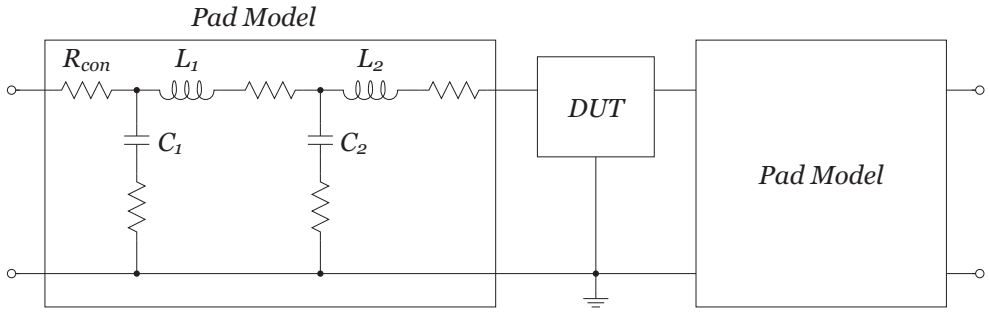


Figure 3.2 Broadband circuit model for the test fixture surrounding the DUT.

3.3 Probe Pad Design

The design of the probe pad and interconnect on lossy substrates involves trade-offs between parasitic capacitance, resistive losses, and port-to-port isolation [43]. Smaller pad parasitics are more easily de-embedded, so pads should have small dimensions and be implemented using the top metal layers in a multi-layer CMOS technology [45]. The minimum pad size is limited by the required probe contact area and foundry design rules. The mechanical integrity of the pad must be maintained, and pads implemented using the top two metal layers in a 6-metal layer CMOS process have been used in this work.

Although using a metal shield underneath the pads substantially increases the parasitic capacitance over an unshielded pad [45], using the bottom layer metal to shield the pads from the lossy silicon substrate has many benefits [43]. A low-noise probe pad and interconnect is desired when characterizing the noise figure performance of small devices. The shielded pad contributes little noise since it does not suffer from the large resistive losses caused by signal coupling to the lossy substrate. A shielded pad also improves port-to-port isolation—the measured isolation is better than 40 dB up to 65

GHz—and minimizes parasitic coupling to the DUT. Connecting the grounds on-chip ensures balanced ground currents and a very low-impedance ground return path. Imbalanced ground currents lead to parasitics that are different from those de-embedded during calibration and can also excite higher-order modes. For these reasons, low-noise shielded pads and interconnects were used for all device characterization.

3.4 Small-Signal De-embedding

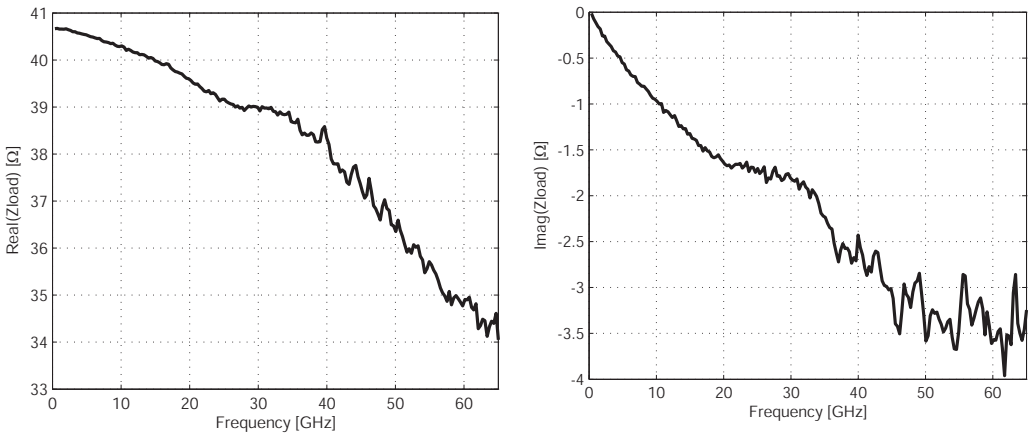
For accurate device characterization at mm-wave frequencies, the effect of the test fixture must be removed using any of a number of techniques such as conventional on-wafer calibration, pad parasitic removal, pad modeling and extraction, or electromagnetic modeling. Pad de-embedding involves the use of on-wafer dummy test structures to characterize the pad parasitics. Since material properties like conductor and dielectric thickness and substrate resistivity vary across the wafer, it is best to put the dummy test structures close to the DUT. Several copies of the dummy structures across the chip can be used for improved de-embedding accuracy. In this section, the common techniques used to de-embed the pad and interconnect are reviewed. The advantages and drawbacks of each for de-embedding on lossy substrates at mm-wave frequencies are highlighted.

3.4.1 SOLT calibration

The most commonly used method to calibrate VNAs is SOLT calibration. This requires four on-wafer standards: a short, an open, a 50-Ω load resistance, and a 50-Ω transmission line of known electrical length. These standards are most easily implemented on an insulating substrate (such as alumina or GaAs) where it is possible to realize accurate passives with low parasitics.

Additionally, the 50-Ω loads are often laser-trimmed. If the electrical characteristics of the standards are known *a priori*, as is the case for the ISS, then SOLT calibration can be used well into the mm-wave frequency regime.

Realizing a precise broadband resistive load presents the greatest difficulty for SOLT standards fabricated on a conductive silicon substrate. The resistor is typically implemented using polysilicon, which has very large parasitics to the substrate due to its close proximity. Even though the size of the resistor is small (<12 μm), there is still a large frequency-dependence in the impedance of the load as shown in Fig. 3.3. Given the inability to fabricate precise broadband loads on silicon, SOLT calibration is not a viable option for accurate mm-wave de-embedding on CMOS.



(a) (b)

Figure 3.3 (a) Resistance and (b) reactance of a polysilicon resistor in CMOS.

3.4.2 TRL calibration

To avoid the problem of fabricating broadband well-characterized loads, the through-reflect-line (TRL) calibration technique was developed [46]. Only three standards are necessary: a zero-length through, a standard with high reflection coefficient (typically either a short or open), and a transmission line

(Fig. 3.4). The reflect standard need not be well-characterized, as only its phase must be known to select between two numerically valid solutions. Therefore, TRL does not suffer from the parasitic fringing capacitance and inductance of the dummy test structures that limit the accuracy of the open and open-short de-embedding techniques described later in Sections 3.4.3 and 3.4.4 [47]. The most important advantage of TRL over SOLT calibration is that well-characterized transmission lines are more easily fabricated on-wafer than precise resistive loads.

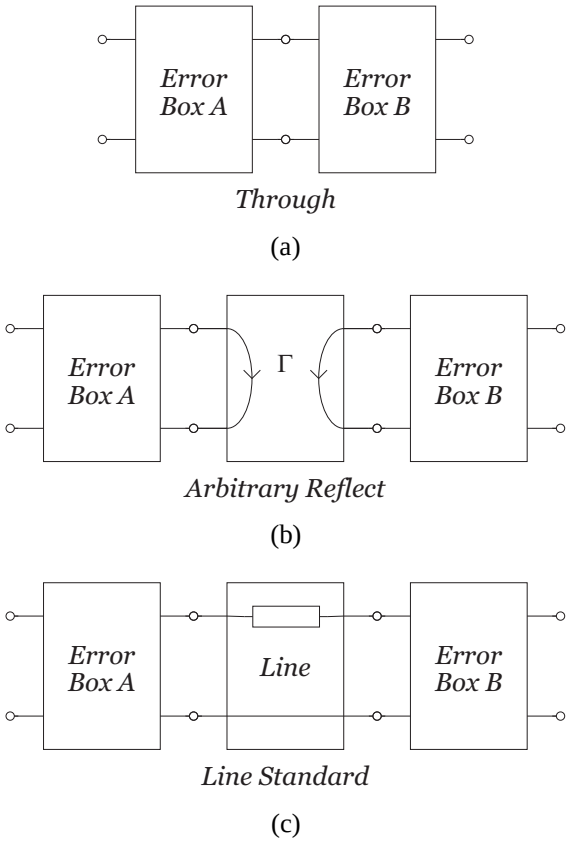


Figure 3.4 TRL calibration structures. (a) Through (b) Reflection (c) Line.

There are several disadvantages associated with TRL calibration. First, for accurate de-embedding, the line standard needs to have electrical length between 20° and 160°. Therefore, TRL calibration typically can only be

applied over a bandwidth no larger than 8:1 [48] and requires many lines of varying lengths for broadband de-embedding. This makes multiline TRL calibration cumbersome and tedious. Furthermore, for low frequencies, the line standard becomes prohibitively long and consumes a significant amount of die area.

The more fundamental problem with TRL calibration is the fact that the de-embedded *S*-parameters are referenced to the complex-valued Z_0 of the line. As was described in Section 2.5.1, the characteristic impedance of a transmission line can be expressed as

$$Z_0 \approx Z\left(1+\frac{j}{2}\left(\frac{1}{Q_C}-\frac{1}{Q_L}\right)\right) \qquad \textbf{(3.1)}$$

For lines fabricated on silicon (and even for lines fabricated on insulating substrates), there is a large variation in Q_L and Q_C with frequency (Fig. 2.19). This has an important impact on the phase of the resulting *S*-parameters. Even with the advances of 3-D electromagnetic (EM) simulators, the complex characteristic impedance of on-wafer transmission lines can only be simulated with limited precision because of process variations in conductor dimensions and dielectric thickness as well as poorly-characterized material properties such as surface roughness and frequency-dependent loss tangents. There is active research in developing an accurate theoretically-sound measurement-based technique to determine the Z_0 of lines fabricated on lossy silicon substrates [47][49].

3.4.3 Open de-embedding

Since the measurement system has already been calibrated to the probe tips with an ISS, it is simpler to remove the pad parasitics using two-tier de-embedding instead of the conventional calibration techniques described in

Sections 3.4.1 and 3.4.2. The most basic de-embedding strategy is open de-embedding [50]. The dominant pad parasitic, particularly at low frequencies, is the parasitic capacitance from the pad to ground (C_1 and C_2 in Fig. 3.2). For this technique to be valid, the equivalent circuit model that is assumed is shown in Fig. 3.5. Y_1 and Y_2 can be determined by measuring an open structure, and the port-to-port coupling admittance, Y_3, can often be made negligible with adequate spacing between the input and output probe pads. The simplicity of requiring only one open test structure makes this de-embedding approach attractive. The pads are removed by subtracting the *Y*-parameters of the open from the device measurement.

Open de-embedding implicitly assumes that the inductance and resistance of the interconnect between the pad and the DUT are small and can be neglected. This assumption is valid at lower frequencies, but at higher frequencies when the impedances of the series components (L_1 and L_2 in Fig. 3.2) are comparable to the impedance of the probe pad capacitance, the assumptions used for open de-embedding become invalid. Open de-embedding is generally adequate for frequencies up to only a few GHz.

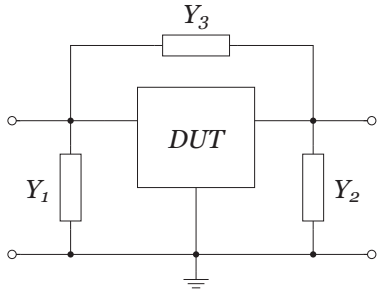


Figure 3.5 Equivalent circuit model used for open de-embedding.

3.4.4 Open-short de-embedding

With the addition of one dummy short structure, it is possible to also remove the series impedances between the pad and the DUT [51]. The equivalent circuit model assumed for open-short de-embedding is shown in Fig. 3.6. The measurement of the open captures the shunt parasitics, while the short is the series impedance in parallel with the shunt parasitics. For best results, the interconnect between the pads and DUT should be kept as short as possible and wide metal conductors should be used to minimize the series impedance. Removal of the pads is performed by subtraction of the *Y*- and *Z*-parameters of the open and short from the embedded DUT measurement. Open-short de-embedding is a significant improvement over simple open de-embedding, and is the technique most commonly used for device characterizations.

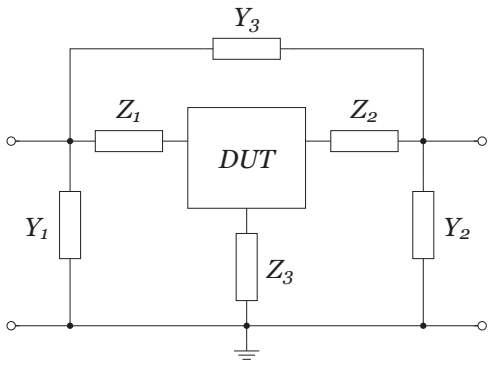


Figure 3.6 Equivalent circuit model used for open-short de-embedding.

At millimeter-wave frequencies, the true distributed nature of the pads (Fig. 3.2) makes the lumped model assumption of Fig. 3.6 inaccurate. Furthermore, the dummy test structures have some parasitics that are not present when the DUT is connected. The dummy open has extra parasitic fringing capacitance, while the dummy short has extra parasitic inductance to ground. This leads to over-de-embedding and optimistic device performance [47]. The exact frequency that open-short de-embedding becomes inaccurate is

a strong function of the DUT being characterized. A large DUT can be characterized more precisely to higher frequencies than a minimum-sized device.

3.4.5 Other de-embedding techniques

Other de-embedding methods have been proposed to solve some of the limitations of the de-embedding techniques described above. The procedure described by Kolding [52] is based on a more complex model for the fixture, and requires more dummy test structures in order to characterize the pad parasitics. Unfortunately, this suffers from a similar problem with open-short de-embedding in that parasitics in the de-embedding structures, which are not present once the DUT is inserted, will lead to over-de-embedding. Four-port de-embedding [53] has also been proposed, but assumes the ability to realize a frequency-independent (although not required to be 50 Ω) load, which is not a valid assumption for silicon-based standards as was shown in Fig. 3.3.

3.4.6 Summary

In this section, several calibration and de-embedding techniques have been described. The first step is to use standard calibration substrates provided by the probe manufacturer to set the reference plane at the probe tip. SOLT and TRL calibration to move the reference plane to the input and output of the DUT are not appropriate for on-silicon de-embedding because they are founded on the ability to fabricate a precise load resistor and transmission line, respectively. These are not valid assumptions for CMOS measurements. Open-short de-embedding is described that is based on a simple fixture model and requires only two de-embedding structures—one open and one short. It will be shown in the next section that open-short de-embedding provides good

accuracy for obtaining the small-signal parameters of both transistors and transmission lines to 65 GHz. For this work, open-short de-embedding is used exclusively.

3.5 Small-Signal De-embedding Verification

Several de-embedding techniques were described in the previous section along with their associated problems. Although it is clear that none of the mentioned de-embedding methods can perfectly remove the effects of the fixture to arbitrarily high frequencies, it is also impossible to determine exactly how much error the de-embedding process introduces. (If the error could be determined exactly, then the verification technique could also be used to de-embed perfectly.) However, there are some verification metrics that can provide a good indication of the validity range of the de-embedding. These tests can only determine when the de-embedding becomes *invalid*, rather than ensuring the de-embedding is valid. However, they provide a good foundation for comparing various pad structures and also understanding the effect of the DUT on the de-embedding accuracy.

In this section, two verification tests are described. The first, and most straightforward, test is to de-embed a zero-length through. In Section 3.5.2, a new test involving the use of TRL calibration and Z_0-invariant parameters for verification is introduced and developed. This test is particularly useful when the DUT is a transistor or transmission line, in which case several important circuit parameters can be accurately determined to very high frequencies. The theory and derivations behind the TRL verification will be described.

3.5.1 Through verification

The simplest verification is to de-embed a zero-length through. This test is

especially useful for open and open-short de-embedding, as the through is not used as part of the de-embedding standard. The *S*-parameters for the through, after de-embedding, are ideally,

$$S_{thru} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{3.2}$$

A plot of S_{11} for both open and open-short de-embedding is shown in Fig. 3.7. From (3.2), it can be seen that perfect de-embedding would result in $S_{11} = -\infty$ dB. As expected, the result from the open de-embedding begins to deviate from perfect de-embedding at a much lower frequency than for open-short de-embedding. The result from open-short de-embedding is better at 65 GHz than from open de-embedding at 7 GHz.

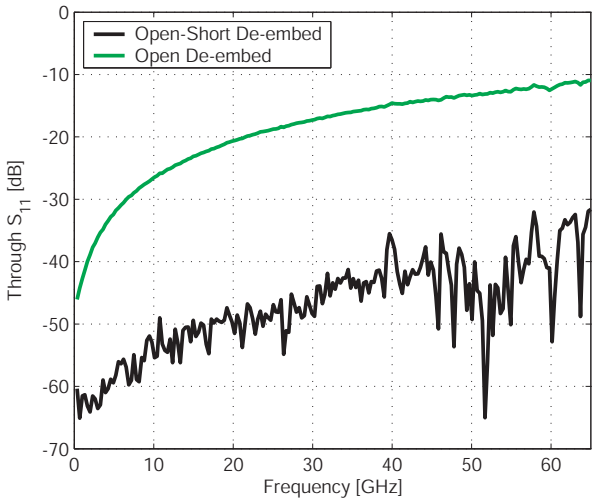


Figure 3.7 Through verification for open and open-short de-embedding.

3.5.2 TRL verification

As described in Section 3.4.2, the results after TRL calibration are *S*-parameters of the DUT referenced to the complex Z_0 of the transmission line standard. As mentioned, the characteristic impedance of the line is not easily

determined; however, several parameters of the DUT that are *invariant* to the complex Z_0 of the line standard can be computed. Therefore, these parameters of the intrinsic DUT can be determined accurately using TRL calibration well into the mm-wave frequency range, even without knowledge of the Z_0 of the line. Furthermore, since the DUT parameters are also independent of the propagation constant of the line standard, *any* dummy test structure that is symmetric ($z_{11} = z_{22}$) and reciprocal ($z_{12} = z_{21}$) can be used as the "line" standard to compute these parameters.

Consider the formula relating the *Z*- and *Y*-parameters to *S*-parameters referenced to a complex Z_0 [54],

$$[z] = Z_0 ([\mathbf{1}] + [s])([\mathbf{1}] - [s])^{-1} \quad \textbf{(3.3)}$$

$$[y] = \frac{1}{Z_0}([\mathbf{1}] - [s])([\mathbf{1}] + [s])^{-1} \quad \textbf{(3.4)}$$

where $[\mathbf{1}]$ is the unit diagonal matrix. From (3.3) and (3.4), it is obvious that Z_0 acts as a scale factor for the *Z*- and *Y*-parameters. Thus, by taking the appropriate ratios or products of the *Z*- and *Y*-parameter elements, these values will be Z_0-invariant. This provides up to three independent quantities if the DUT is neither symmetric nor reciprocal (e.g., a transistor). These quantities can be determined precisely using TRL calibration and can be used to compare against the values resulting from open or open-short de-embedding to verify the accuracy of these de-embedding techniques.

3.5.3 Z_0-invariant parameters for transistors

Consider taking ratios of the *Y*-parameters. This results in three well-known and commonly-used parameters that characterize transistor performance:

$$h_{21} = y_{21} / y_{11} \quad \textbf{(3.5)}$$

$$MSG = y_{21}/y_{12} \tag{3.6}$$

$$A_{v0} = -y_{21}/y_{22} \tag{3.7}$$

where h_{21} is the short-circuit current gain, *MSG* is the maximum stable gain, and A_{v0} is the open-circuit voltage gain. Notice from (3.5) that it is possible to accurately characterize the device f_t to very high frequencies using TRL calibration. The most sensitive parameter to de-embedding errors is typically A_{v0}. Comparing the values of h_{21}, *MSG*, and A_{v0} resulting from open or open-short de-embedding with the value from TRL calibration gives a good indication of when the de-embedding error becomes large.

For the data shown in Fig. 3.8, A_{v0} computed after open de-embedding begins to deviate from the TRL verification data at 10–20 GHz for a small 40×1 μm/0.13 μm device. Above 30 GHz, there is a significant discrepancy. Notice that not all parameters exhibit the same sensitivity to de-embedding errors. In particular, *MSG* is extremely robust to de-embedding inaccuracies. Comparing Figs. 3.8 and 3.9, it is clear that open-short de-embedding provides a drastic improvement over open de-embedding. There is a small deviation in A_{v0} above 45–50 GHz, which is consistent with the *S*-parameter data of the de-embedded DUT that show an anomalous "kink" around 45 GHz. Thus, this verification technique provides supporting evidence that the kink is caused by de-embedding errors as opposed to an inherent behavior of the transistor. The data in Fig. 3.10 shows open-short data for a larger 80×1 μm/0.13 μm device, with nearly perfect fitting. Intuitively, one would expect smaller de-embedding errors for larger devices. This data confirms that intuition, and suggests that for large devices, open-short de-embedding can be used to accurately de-embed up to 65 GHz.

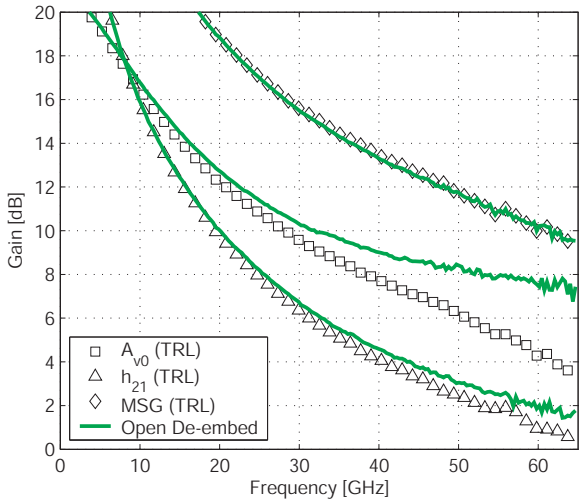


Figure 3.8 Plot of the TRL verification parameters against those computed after open de-embedding for a 40×1 µm/0.13 µm NMOS transistor.

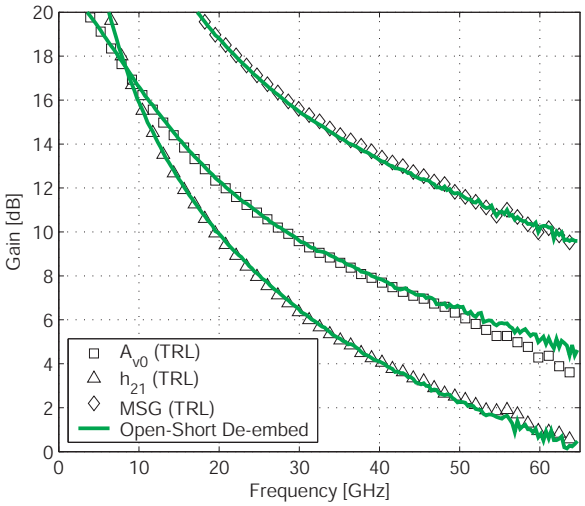


Figure 3.9 Plot of the TRL verification parameters against those computed after open-short de-embedding for a 40×1 µm/0.13 µm NMOS transistor.

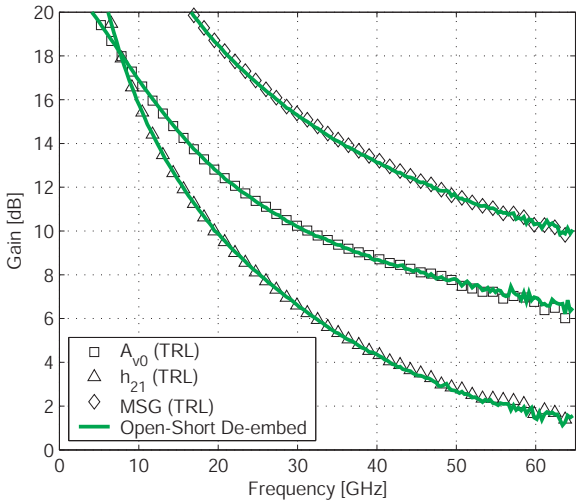


Figure 3.10 Plot of the TRL verification parameters against those computed after open-short de-embedding for an 80×1 μm/0.13 μm NMOS transistor.

3.5.4 Z_0-invariant parameters for transmission lines

If the DUT is a lossy transmission line with characteristic impedance Z_L and propagation constant γ, the *ABCD*-parameters are [55]

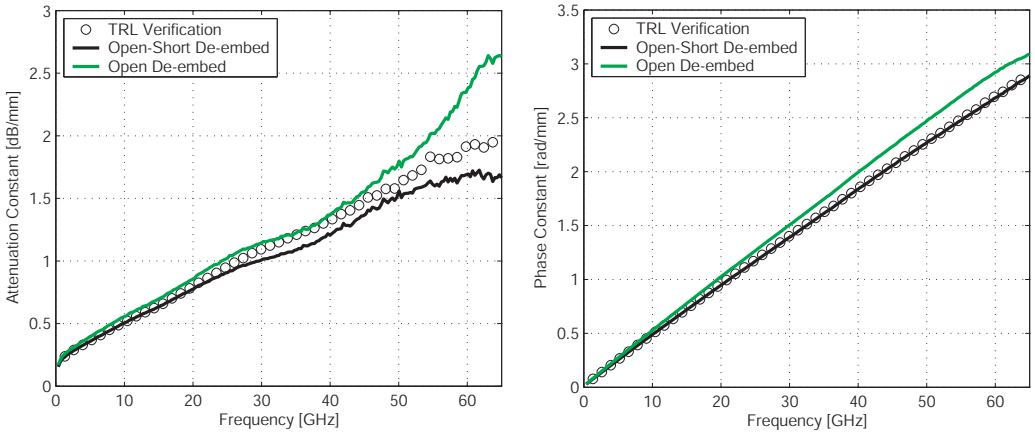
$$[ABCD] = \begin{bmatrix} \cosh \gamma l & Z_L \sinh \gamma l \\ \dfrac{\sinh \gamma l}{Z_L} & \cosh \gamma l \end{bmatrix} \tag{3.8}$$

Since transmission lines are symmetric and reciprocal, this reduces the number of Z_0-invariant parameters to one. Relating the elements of the *ABCD*-parameters to *Z*-parameters [17],

$$A = \frac{z_{11}}{z_{21}} = \cosh \gamma l \tag{3.9}$$

So, the complex propagation constant, γ, of a transmission line device is independent of the Z_0 of the line standard and can be determined accurately to very high frequencies. From Fig. 3.11, it appears that both open and open-short

de-embedding provide good performance up to 40–50 GHz. However, as mentioned earlier, the TRL verification procedure only checks for *invalid* de-embedding, and good fitting does not guarantee accurate de-embedding. The extracted value for the characteristic impedance of the coplanar waveguide line from open and open-short de-embedding is shown in Fig. 3.12. The erratic data above 20 GHz for open de-embedding is more consistent with the results from the verification tests performed in Sections 3.5.1 and 3.5.2. Therefore, it is best to examine the data from all of the de-embedding verification tests to understand the approximate useful range for the de-embedding approach. This example serves to highlight the challenges involved in distinguishing between de-embedding inaccuracies and the actual electrical characteristics of the DUT.



(a) (b)

Figure 3.11 (a) Attenuation and (b) phase constants for a CPW transmission line extracted after open and open-short de-embedding.

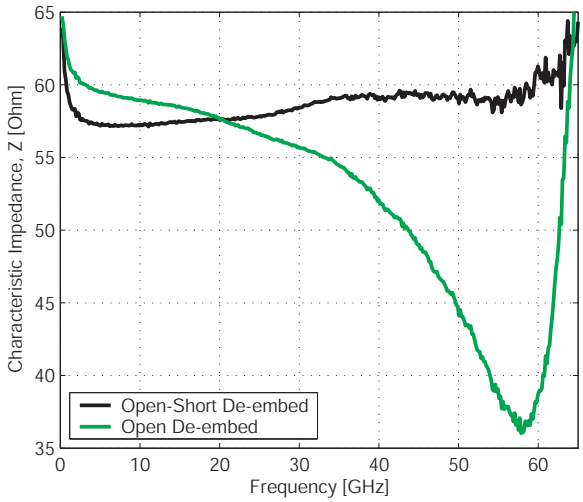


Figure 3.12 Characteristic impedance for a CPW transmission line extracted after open and open-short de-embedding.

3.5.5 Summary

A key challenge for comparing de-embedding techniques is to determine when the actual measurement deviates from the model used for de-embedding. No technique exists that can perfectly de-embed to arbitrarily high frequencies. Two methods to quantify inaccuracies are presented: through verification and TRL verification. These metrics can confirm that a given de-embedding method is inaccurate, but does not guarantee that a de-embedding method is accurate. However, they still provide good insight into which technique is better. Using these derived metrics, it is found that open-short de-embedding is sufficiently accurate for small-signal de-embedding of the probe pads from the device up to 65 GHz for both transistors and transmission lines.

3.6 Device Noise De-embedding

In addition to de-embedding the probe pads to obtain the *S*-parameters of the

intrinsic DUT, it is also critical to remove the effect of the fixture from on-wafer noise characterization. The probe pads impact the noise measurements of the DUT in two ways. First, the losses in the probe pads will contribute thermal noise to the measurement. Second, the noise parameters of the DUT will be transformed as it goes through the reactive network of the pads. These effects must be compensated in order to obtain the intrinsic DUT noise parameters that will be used for device noise modeling. This section describes the theory of noisy two-ports and presents a step-by-step method for probe pad noise de-embedding.

3.6.1 Noisy two-port theory

Any noisy linear two-port network with internal noise sources can be represented using an equivalent noiseless two-port with two external noise generators [56]. The noiseless two-port is identical to the original noisy two-port with the exception that the internal noise sources have been removed. Different equivalent representations for the noisy two-port model are shown in Fig. 3.13. It should be noted that, in the most general case, the two noise sources are correlated.

Although all three noise representations are equivalent, the different models are used based on computational convenience or to provide additional intuitive understanding. If the two-port is represented using the admittance (*Y*) matrix then Fig. 3.13a is the natural choice. If the two-port is represented using the impedance (*Z*) matrix then Fig. 3.13b is the most convenient. When using the chain (*ABCD*) matrix, the representation in Fig. 3.13c is preferred. As a practical example, the *PRC* noise model for the intrinsic transistor introduced in Section 2.4.2 models the noise with two external noise current sources at the input and output. Another common example is the model used for noise figure analysis. Since noise figure is not dependent on the transfer function through

an arbitrary noise-free two-port, the representation in Fig. 3.13c that specifies the noise as two input noise sources is typically used.

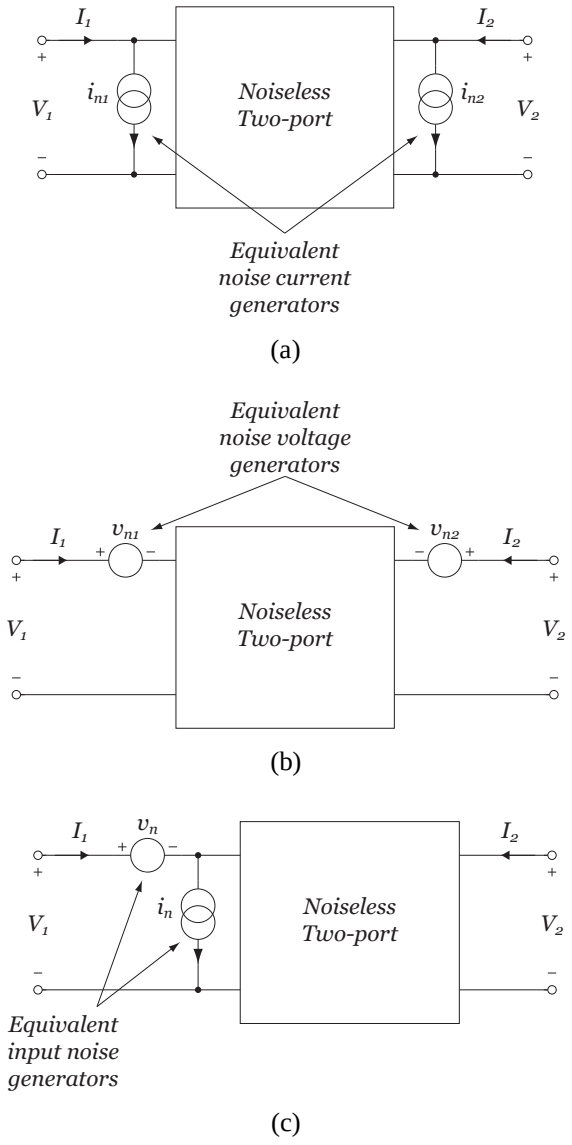


Figure 3.13 Equivalent circuit of noisy two-port with (a) input and output noise current sources i_{n1} and i_{n2}, (b) input and output noise voltage sources v_{n1} and v_{n2}, and (c) noise voltage source v_n and noise current source i_n at input.

3.6.2 Noise correlation matrices

The theory of noise correlation matrices is analyzed in detail in [57], and only the results relevant to noise de-embedding are summarized here. For the equivalent noise representations shown in Fig. 3.13, the noise correlation matrices are defined as

$$C_Y = \frac{1}{4kT\Delta f}\begin{bmatrix} \overline{i_{n1}i_{n1}^*} & \overline{i_{n1}i_{n2}^*} \\ \overline{i_{n2}i_{n1}^*} & \overline{i_{n2}i_{n2}^*} \end{bmatrix} = \frac{1}{4kT\Delta f}\overline{\begin{bmatrix} i_{n1} \\ i_{n2} \end{bmatrix} \cdot \begin{bmatrix} i_{n1} \\ i_{n2} \end{bmatrix}^H} \qquad \textbf{(3.10)}$$

$$C_Z = \frac{1}{4kT\Delta f}\begin{bmatrix} \overline{v_{n1}v_{n1}^*} & \overline{v_{n1}v_{n2}^*} \\ \overline{v_{n2}v_{n1}^*} & \overline{v_{n2}v_{n2}^*} \end{bmatrix} = \frac{1}{4kT\Delta f}\overline{\begin{bmatrix} v_{n1} \\ v_{n2} \end{bmatrix} \cdot \begin{bmatrix} v_{n1} \\ v_{n2} \end{bmatrix}^H} \qquad \textbf{(3.11)}$$

$$C_A = \frac{1}{4kT\Delta f}\begin{bmatrix} \overline{v_n v_n^*} & \overline{v_n i_n^*} \\ \overline{v_n^* i_n} & \overline{i_n i_n^*} \end{bmatrix} = \frac{1}{4kT\Delta f}\overline{\begin{bmatrix} v_n \\ i_n \end{bmatrix} \cdot \begin{bmatrix} v_n \\ i_n \end{bmatrix}^H} \qquad \textbf{(3.12)}$$

The noise correlation matrix fully captures the behavior of the noisy two-port. The noise sources in (3.10)–(3.12) are defined by their single-sided noise spectral density (positive frequencies), since those are most commonly used by circuit designers. This introduces a factor of 2 between these equations and those presented in [57]. Additionally, the parameter $4kT$ has been factored out for notational simplification.

A practical example where the noise correlation matrix is known is for a lossy passive network. It can be shown, based on thermodynamic principles, that the noise correlation matrix is [58]

$$C_Y = \frac{1}{2}\left(Y + Y^H\right) = \mathrm{Re}\{Y\} \qquad \textbf{(3.13)}$$

$$C_Z = \frac{1}{2}\left(Z + Z^H\right) = \mathrm{Re}\{Z\} \tag{3.14}$$

In order to transform between different noise representations, a transformation matrix is calculated. For example, to convert between the impedance and chain representations,

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} v_n \\ i_n \end{bmatrix} + \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} V_2 \\ -I_2 \end{bmatrix} \tag{3.15}$$

$$\begin{bmatrix} V_1 \\ I_1 \end{bmatrix} = \begin{bmatrix} v_{n1} \\ 0 \end{bmatrix} + \begin{bmatrix} A & B \\ C & D \end{bmatrix} \left(\begin{bmatrix} V_2 \\ -I_2 \end{bmatrix} - \begin{bmatrix} v_{n2} \\ 0 \end{bmatrix} \right) \tag{3.16}$$

$$\begin{bmatrix} v_n \\ i_n \end{bmatrix} = \begin{bmatrix} 1 & -A \\ 0 & -C \end{bmatrix} \begin{bmatrix} v_{n1} \\ v_{n2} \end{bmatrix} = T_{AZ} \begin{bmatrix} v_{n1} \\ v_{n2} \end{bmatrix} \tag{3.17}$$

where

$$T_{AZ} = \begin{bmatrix} 1 & -A \\ 0 & -C \end{bmatrix} \tag{3.18}$$

Using (3.10)–(3.12), the transformation is

$$C_A = T_{AZ} C_Z T_{AZ}^H \tag{3.19}$$

When cascading two noisy networks represented by their chain matrices (A_1, A_2) and noise correlation matrices (C_{A1}, C_{A2}), the resulting network is

$$A_{tot} = A_1 A_2 \tag{3.20}$$

$$C_{Atot} = C_{A1} + A_1 C_{A2} A_1^H \tag{3.21}$$

If A_1 is a lossy pad to be de-embedded from the cascaded network of the pad with the intrinsic device,

$$C_{A2} = \left(A_1\right)^{-1} \left(C_{Atot} - C_{A1} \right) \left(A_1^H\right)^{-1} \tag{3.22}$$

3.6.3 Noise correlation matrices and noise parameters

A common way to specify the noise performance of a device is by the minimum noise figure (F_{min}), noise resistance (R_n), and optimal source admittance (Y_{opt}) [56]. The unique mapping from noise parameters to correlation matrix is

$$C_A = \begin{bmatrix} R_n & \frac{F_{\min}-1}{2} - R_n Y_{opt}^* \\ \frac{F_{\min}-1}{2} - R_n Y_{opt} & R_n \left|Y_{opt}\right|^2 \end{bmatrix} \quad \textbf{(3.23)}$$

To determine the noise parameters from the correlation matrix,

$$R_n = C_{A11} \quad \textbf{(3.24)}$$

$$Y_{opt} = \sqrt{\frac{C_{A22}}{C_{A11}} - \left[\mathrm{Im}\left(\frac{C_{A12}}{C_{A11}}\right)\right]^2} + j\,\mathrm{Im}\left(\frac{C_{A12}}{C_{A11}}\right) \quad \textbf{(3.25)}$$

$$F_{\min} = 1 + 2\left(C_{A12} + C_{A11} Y_{opt}^*\right) \quad \textbf{(3.26)}$$

3.6.4 Noise de-embedding of probe pads

Based on the derivation of the noise correlation matrix described in Sections 3.6.2 and 3.6.3, the following methodology can be used to obtain the noise parameters of an intrinsic device when only the noise parameters of the DUT embedded within the probe pads are available.

- Determine the two-port model for the pad and represent this as an impedance matrix (Z_{pad}) and chain matrix (A_{pad}).
- Using (3.14), (3.17), and (3.19), determine the noise correlation matrix C_{Apad}.
- Using (3.23), convert the embedded noise parameters to a correlation

matrix (C_{Atot}).

- Given A_{pad}, C_{Apad}, and C_{Atot}, use (3.22) to compute the noise correlation matrix of the DUT (C_{Adut}).
- Compute the DUT noise parameters from the DUT noise correlation matrix using (3.24)–(3.26).

Using this methodology, the loss and impedance transformation of the wafer probe pads can be de-embedded from the noise parameters of the embedded device to yield the noise parameters of the intrinsic device. This technique was applied to the measured data described in Section 3.6.5 to yield the device noise parameters shown in Figs. 2.11 and 2.12.

3.6.5 On-wafer noise parameter measurements

The noise parameters have been measured for four CMOS devices—three common-source NMOS transistors, one cascode structure—using the setup and methodology described in [59]. The schematic of the setup is shown in Fig. 3.14 and a photograph of the system is displayed in Fig. 3.15. The setup used for on-wafer noise parameter measurements consists of a *V*-band noise source, waveguide tuner, low noise receiver, and automated noise figure meter.

The noise figure varies with different source reflection coefficients Γ_S based on the well-known relationship [56]

$$F = F_{\min} + 4r_n \frac{\Gamma_S - \Gamma_{opt}^2}{\left(1 - |\Gamma_S|^2\right)|1 + \Gamma_{opt}|^2} \qquad \textbf{(3.27)}$$

where F_{min} is the minimum noise figure, r_n is the noise resistance normalized to 50 Ω, and Γ_{opt} is the optimal source reflection coefficient for minimum noise figure. As the source impedance is varied using the calibrated tuner, the noise figure is measured. In theory, only four measurements are necessary; however, in practice, given the errors associated with mm-wave noise measurements and

uncertainties in the characteristics of the tuner, additional points are measured and the noise parameters are determined by fitting (3.27) to the data.

Through careful calibration, the noise parameters can be determined with the reference plane at the probe tips. The probe pads were then de-embedded in order to obtain the noise parameters of the intrinsic device.

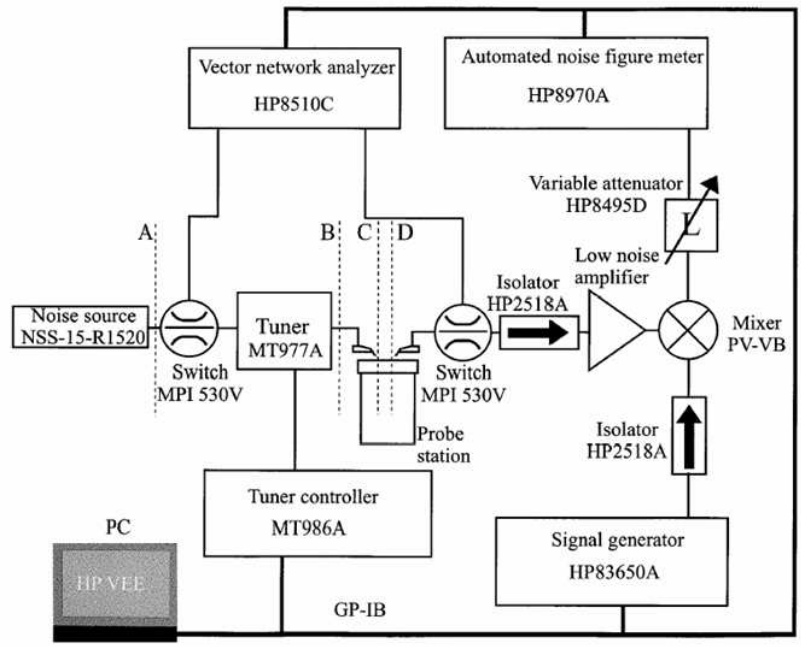


Figure 3.14 Noise parameter measurement system.

3.6.6 Summary

Transistor noise limits the noise figure and phase noise of CMOS receivers. In this section, the theory behind noisy two-ports is described and a procedure used to de-embed the impact of the probe pads on the device noise parameters is presented. The only assumption is that the *S*-parameters of the passive probe pad is accurately known. This model is implicitly known by the open-short de-embedding technique described in Section 3.4.4. Using this procedure, the noise parameters of the intrinsic DUT can be determined and be used for

small-signal transistor noise model extraction up to mm-wave frequencies, as described in Section 2.4.2.

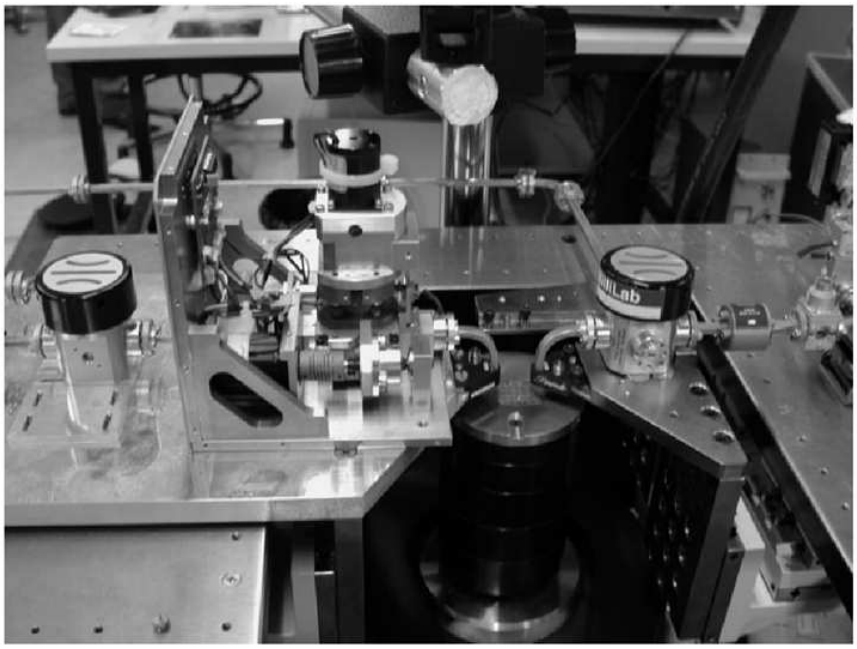


Figure 3.15 Photograph of the on-wafer noise parameter measurement system.

3.7 Conclusion

The ability to accurately characterize the *S*-parameters and noise parameters of an intrinsic DUT is critical for good model extraction. De-embedding errors will result in models that are fit to invalid data. Ensuring valid data for the intrinsic device using on-wafer measurements allows models to be extracted that are accurate at mm-wave frequencies and circuits based on these models to be designed with confidence.

For *S*-parameters, open-short de-embedding has been chosen due to the simplicity of the dummy de-embedding structures and given the fact that good broadband loads and well-defined transmission lines cannot be fabricated on a

lossy silicon substrate. Through verification has been used to determine the accuracy of the fixture model depicted in Fig. 3.6. A new verification technique based on TRL calibration and Z_0-independent parameters has been introduced and was used to verify the validity of open-short de-embedding on both transistors and transmission lines. The results demonstrate that open-short de-embedding can be used accurately to around 45 GHz for most DUTs, and with some large devices (80×1 μm/0.13 μm), open-short de-embedding can provide good performance up to 65 GHz.

The noise parameters for several CMOS transistors have been measured using a custom on-wafer noise parameter measurement system. The theory of noise correlation matrices is introduced and applied to de-embed the effect of the probe pads from the intrinsic DUT up to 75 GHz.

4

Circuit Design and Measurement Results

4.1 Introduction

This chapter describes the practical implementation of the design and modeling methodology introduced in Chapters 2 and 3. Some of the key considerations to design circuits using a CPW transmission lines are detailed. This culminates in the design of two mm-wave CPW CMOS amplifiers, one operating at 40 GHz and one at 60 GHz, whose measured performance can be predicted very accurately using the simulation models described in Chapter 2. This provides validation that a 130-nm bulk CMOS process can be designed to predictably operate at 60 GHz, given careful design and modeling of the core active and passive devices.

4.2 Device Test Chips

A 5 mm × 4 mm 130-nm CMOS device test chip was first fabricated to allow

process characterization and model extraction. The chip micrograph of the test chip is shown in Fig. 4.1. This chip includes NMOS and cascode devices with varying W_F and N_F, transmission lines of different lengths for multi-line TRL calibration, calibration structures for open-short de-embedding, CPW and microstrip transmission lines of different geometries, series and shunt capacitors, and passive filters. Every device used in the amplifier circuits was based on a component that was measured and modeled on this test chip.

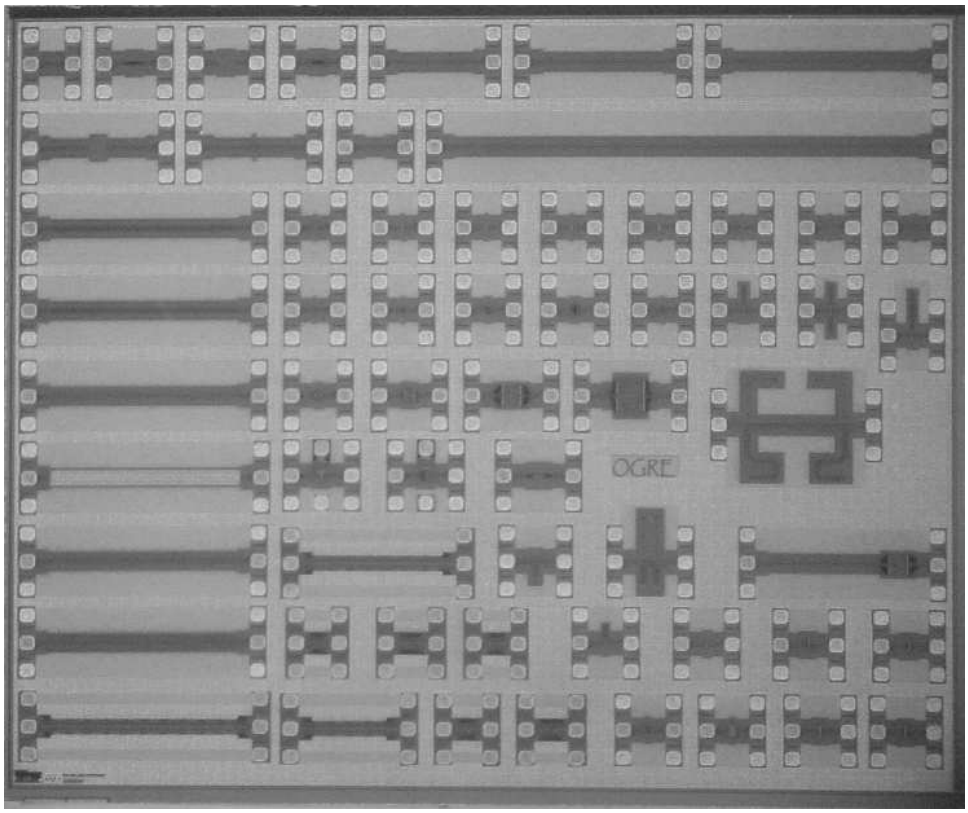


Figure 4.1 Die micrograph of the 130-nm CMOS device test chip.

4.3 CPW transmission line circuit design

As explained in Section 2.5, coplanar waveguide transmission lines on CMOS are preferred to microstrip because of the higher achievable inductive quality

factor. There are some practical considerations when designing circuits using CPW lines.

Copper processing on silicon imposes a peculiar constraint because of the use of chemical mechanical planarization (CMP). This process step requires that the local density of the metal is neither too high nor too low, and can be seen in Fig. 4.2. Ground planes need to be slotted, and dummy metal needs to be added. This also places a constraint on the maximum impedance that can be realized, since the gap cannot be made too large. The design rules regarding the uniformity of metals only gets exacerbated in more advanced technology nodes, and this should be considered early in the design phase. Fortunately, for CPW lines, the majority of the high-frequency current flows on the edge of the metal surrounding the gap, so the slots have only a small effect.

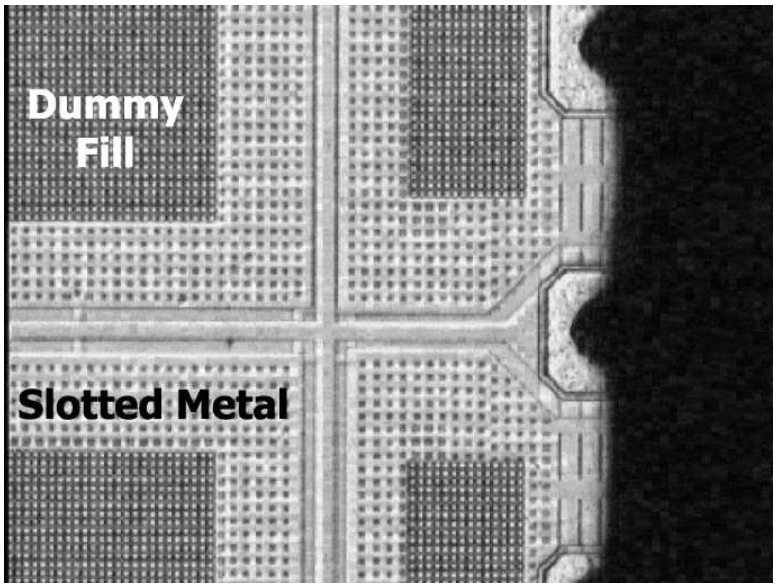


Figure 4.2 CMP constraints for CPW transmission lines on CMOS.

In order to suppress odd-mode propagation at discontinuities in CPW designs, air bridges or underpasses are required. Traditional air bridge technology is accomplished using wirebonds to tie the ground planes together.

Taking advantage of the multi-layer metallization in CMOS, underpasses using a lower level metal can perform the same function. The basic reason that bridges are needed at the discontinuities is to ensure that all ground currents remain balanced. This can be seen in Fig. 4.3, where bridges are added at a cross-junction. Based on Kirchoff's current law (KCL) for the signal conductor,

$$I_1 = I_2 + I_3 + I_4 \qquad \textbf{(4.1)}$$

For a transmission line to behave as expected in normal operation, an implicit assumption is that the ground return current is equal and opposite compared to the signal current. Only by adding the bridges at the junction is the current allowed to redistribute itself to achieve this condition.

An important advantage for using transmission lines compared to lumped passives at 60-GHz, is by having a nearby ground plane surrounding every transmission line, the electromagnetic fields are confined locally and the effects of bends are reduced. A meander line, as shown in Fig. 4.3, can be adequately modeled by a simple straight CPW line of a given length. This is particularly useful for interconnect routing and can be used to help compact the layout.

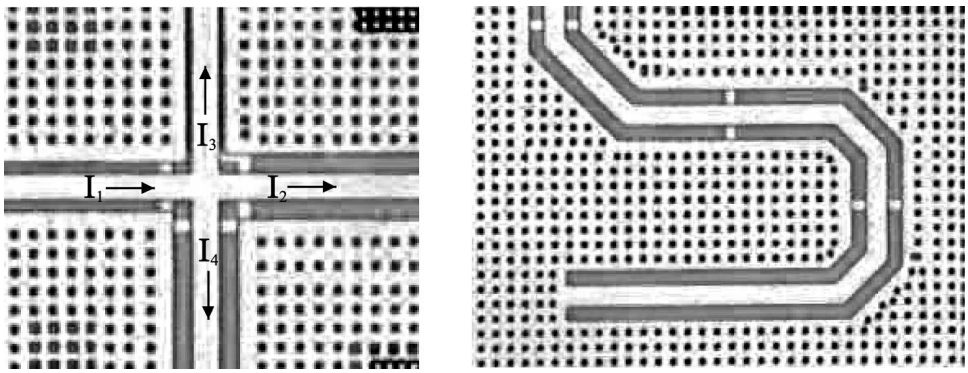


Figure 4.3 Effect of bridges and bends for CMOS CPW lines.

4.4 30-GHz CPW Filter

To validate the electrical passive models, a 30-GHz bandpass filter, composed of series and shunt stubs of the modeled CPW lines, was designed (Fig. 4.4). The topology of the filter is equivalent to a bandpass ladder filter. The shunt resonator is replaced by an open-circuited low-impedance line and a short-circuited high-impedance line. The series transmission line in the center replaces the series resonator. Note that all lines are much shorter than a wavelength ($\lambda = 5$ mm on SiO_2) to minimize loss. Pad models were also extracted from a test chip, and the pads were included as part of the filter. An optimizer was used to fine tune the line lengths.

A die photo of the filter is shown in Fig. 4.5, measuring 0.93 mm × 0.64 mm including pads. Although the transmission lines meander, no special modeling of the bends or junctions was performed, as mentioned in the previous section.

The measured and simulated results for the 30-GHz filter are plotted in Fig. 4.6, demonstrating excellent broadband agreement by using just the simple, scalable electrical models. The measured insertion loss is 2 dB, and the input and output return losses are better than 25 dB. The accurate prediction by the electrical models demonstrates that models for the junctions and bends are not critical, and verifies that there is no significant coupling between the individual lines.

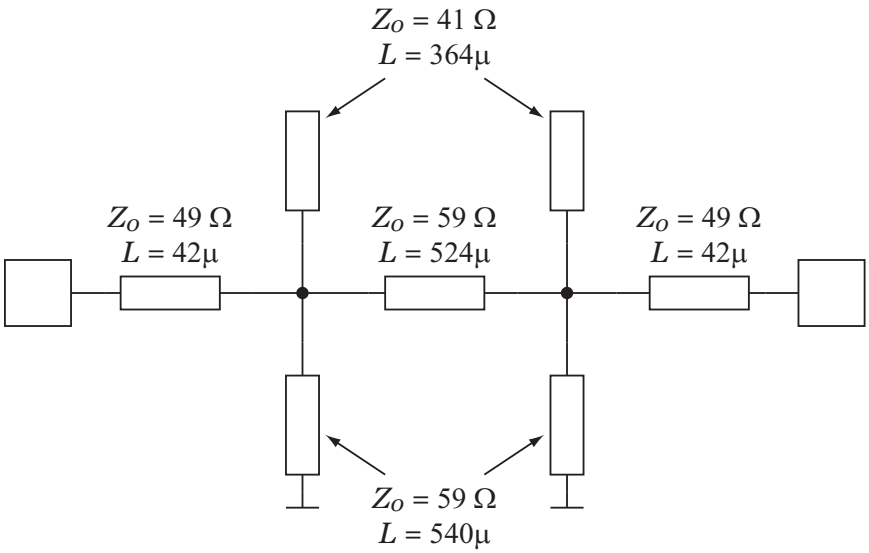


Figure 4.4 Schematic of the 30-GHz CPW filter.

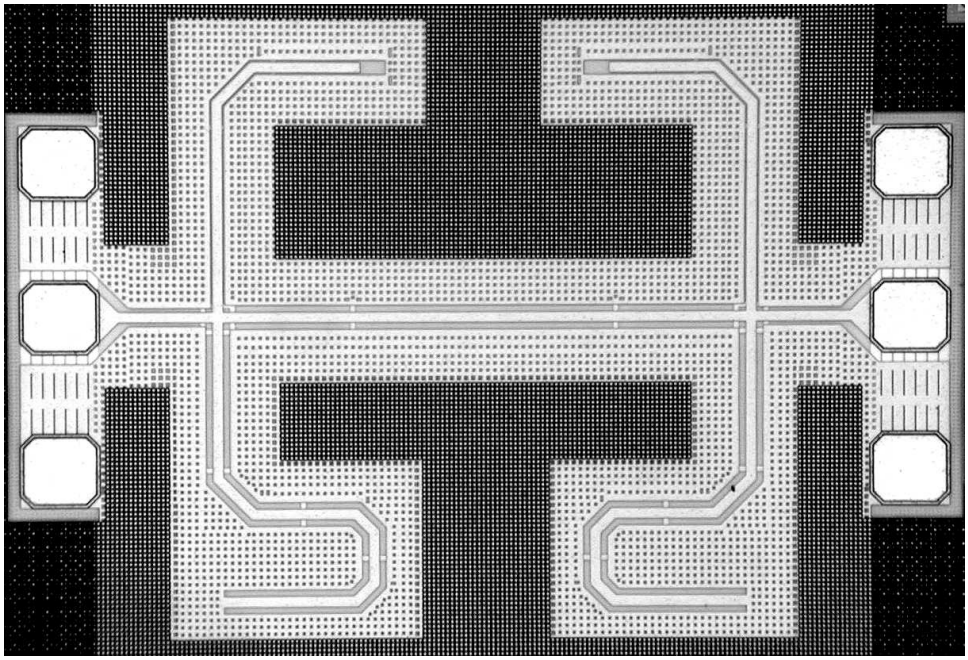


Figure 4.5 Chip micrograph of the 30-GHz CPW filter.

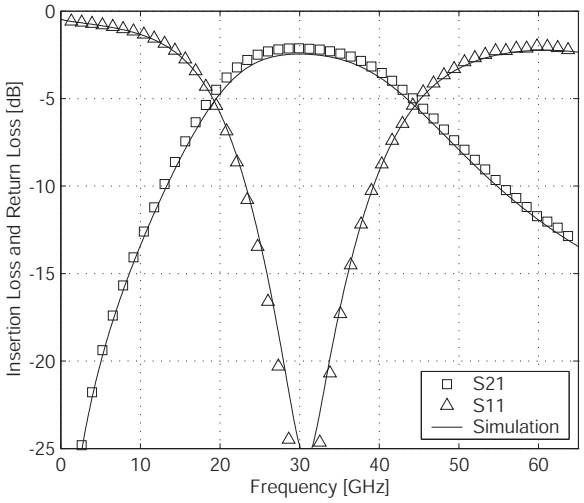


Figure 4.6 Measured and simulated results for the 30-GHz filter.

4.5 40-GHz and 60-GHz Amplifiers

4.5.1 CPW amplifier design

Two wideband mm-wave amplifiers designed as general-purpose amplifiers operating at 40 GHz and 60 GHz have been fabricated in a 130-nm digital CMOS technology with no special analog or RF options. Fig. 4.8 shows the die micrograph of the 40-GHz amplifier, which measures 1.3 mm × 1.1 mm including pads. Fig. 4.9 shows the die micrograph of the 60-GHz amplifier, which measures 1.3 mm × 1.0 mm including pads.

Both amplifiers were designed to have about 25% bandwidth. The topology of the two amplifiers is essentially identical, consisting of three stages of cascode devices with input, output, and interstage reactive matching (Fig. 4.7). The only significant differences between the two amplifiers are the bias currents and lengths of the transmission lines.

Cascode transistors are used in order to reduce the Miller capacitance and improve stability and are unconditionally stable above 27 GHz. The cascode transistors for the 40 GHz amplifier are biased at a current density of 100 μA/μm, with a MAG of 8.9 dB at 40 GHz. For the 60-GHz amplifier, the cascodes are biased at 150 μA/μm, and the MAG is 6.0 dB at 60 GHz. The devices are biased from a V_{DD} of 1.5 V for increased headroom and output power. From simulations, all terminal-pair voltages for the individual transistors remain below the rated breakdown voltages.

CPW transmission lines are used extensively in the design for impedance matching, interconnect wiring, and the bias networks. All lines are kept as short as possible to minimize losses and are significantly shorter than $\lambda/4$ (< 190 μm for the 40-GHz amplifier, < 82 μm for the 60-GHz amplifier). The T-lines at the gate and drain are used to supply bias and are also incorporated into the matching networks. Meander CPW lines are used throughout the 40-GHz design in order to reduce area. The insertion loss of the interstage matching network is 2.5 dB for the 40-GHz design and 1.8 dB for the 60-GHz amplifier. Interestingly, the losses due to the passives are lower at 60 GHz. Although the conductor loss due to skin effect increases with frequency, the lines needed for the matching networks become shorter.

The input and output of the amplifiers are ac-coupled, and the GSG pads are included as part of the design. Both ports are designed to be matched to 50 Ω. The insertion loss of the input matching network is 1.6 dB and 1.3 dB, and the insertion loss of the output matching network is 2.0 dB and 1.6 dB for the 40-GHz and 60-GHz amplifiers, respectively.

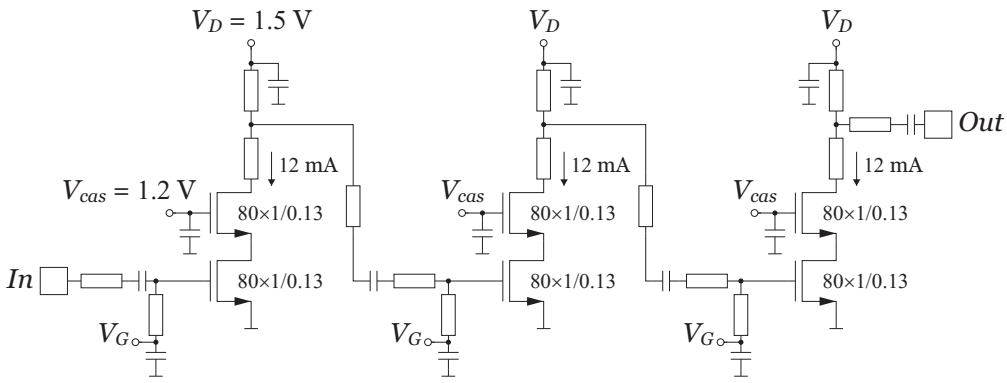


Figure 4.7 Simplified schematic of the 60-GHz 3-stage amplifier using CPW transmission lines.

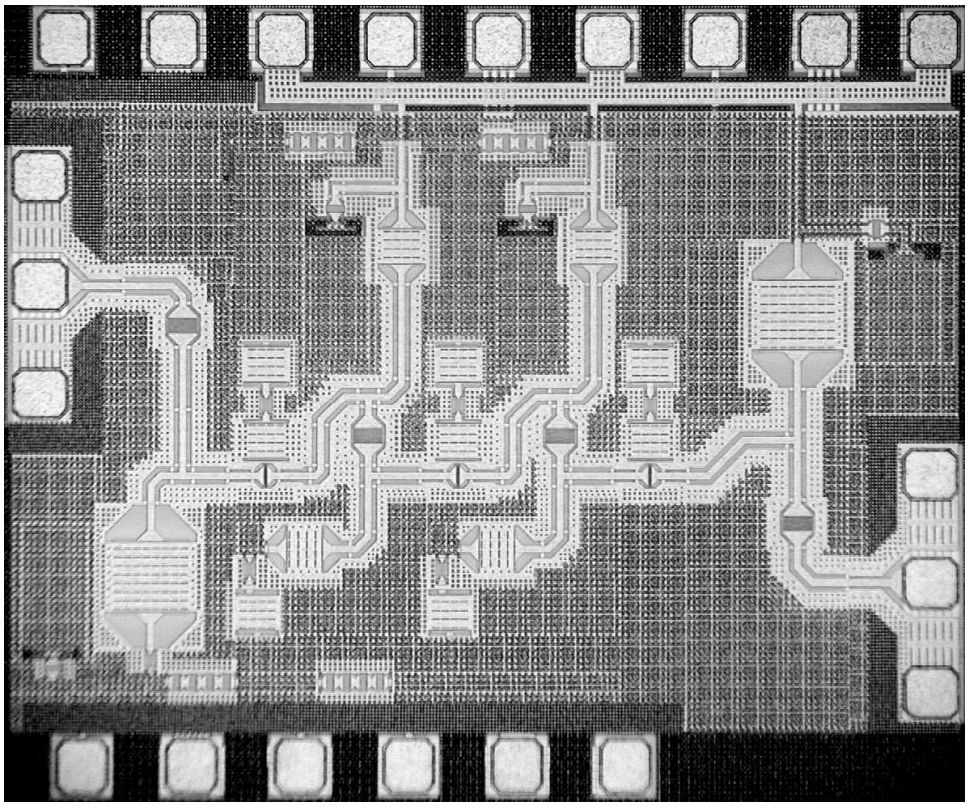


Figure 4.8 Chip micrograph of the 3-stage 40-GHz CPW amplifier.

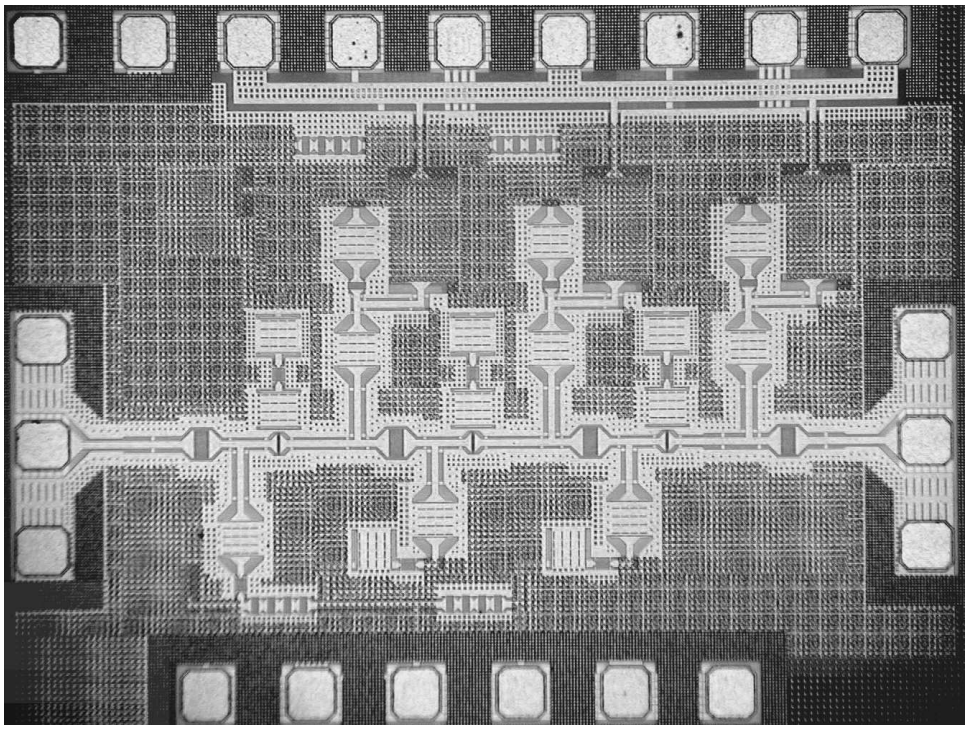


Figure 4.9 Chip micrograph of the 3-stage 60-GHz CPW amplifier.

4.5.2 Measured amplifier results

The measured and modeled *S*-parameters for the 40-GHz amplifier are shown in Fig. 4.10. The amplifier achieves a peak power gain of 19 dB, and input and output return losses are > 15 dB. The 3-dB bandwidth is 34–44 GHz, and the amplifier maintains good return losses across this band. The measured reverse isolation is better than 50 dB up to 65 GHz, indicating that parasitic coupling through the silicon substrate is very small. This isolation is obtained without any special isolation strategy such as deep n-well isolation. The two-tone intermodulation distortion measurements are shown in Fig. 4.12. The measured output 1-dB compression point (P_{1dB}) is −0.9 dBm and IIP3 is −7.4 dBm.

Simulations predict an output P_{1dB} of −1.1 dBm and IIP3 of −10.2 dBm. The noise figure of this amplifier has not been measured, but the simulations show a NF of 5.4 dB. This amplifier dissipates 24 mA from a 1.5-V supply.

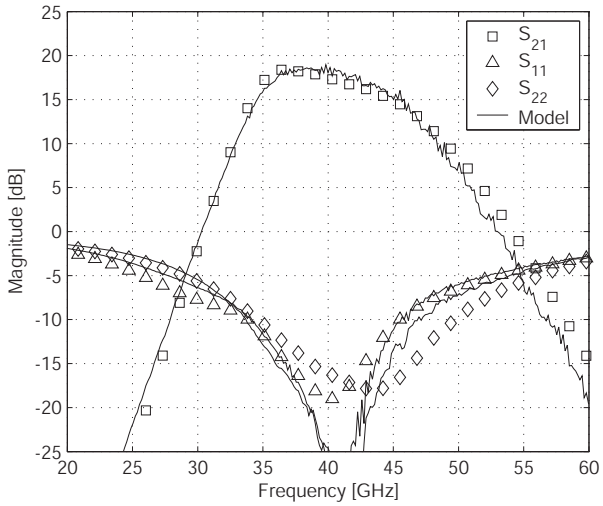


Figure 4.10 Measured (markers) and simulated (lines) S-parameters for the 40-GHz amplifier.

The 60-GHz amplifier has also been characterized, and the *S*-parameters are shown in Fig. 4.11. The amplifier achieves a peak power gain of 12 dB, input and output return losses > 15 dB, and the 3-dB bandwidth is 51–65 GHz. The measured reverse isolation is better than 45 dB up to 65 GHz. The measured output P_{1dB} over frequency is given in Fig. 4.13. The frequency range is limited by the ability for the VNA to drive the amplifier into saturation. At 60 GHz, the measured output P1dB is +2.0 dBm, while simulations predict output P_{1dB} of +1.0 dBm and IIP3 of −0.5 dBm. The output power of the 60-GHz amplifier is higher than the 40-GHz amplifier primarily because of the extra bias current. If the efficiency is kept constant, increasing the current by 50% results in a 3.5 dB increase in output power.

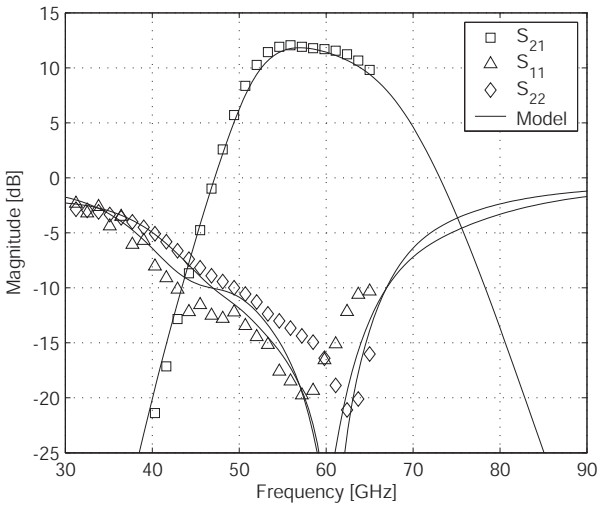


Figure 4.11 Measured (markers) and simulated (lines) S-parameters for the 60-GHz amplifier.

The measured NF of the 60-GHz amplifier is shown in Fig. 4.14. No de-embedding of any on-chip losses was performed since the pads and input match were part of the design. The NF is 8.8 dB at 60 GHz and remains below 9.3 dB up to 63 GHz. Using the default BSIM3 noise model, the simulations predict a noise figure of 6.9 dB. This is because the default BSIM3 noise model does not properly account for effects such as excess short-channel thermal noise or induced gate noise. Using the Pospieszalski noise model described in Section 2.4.2, the simulated noise figure is 9.3 dB at 60 GHz. This demonstrates that the default BSIM noise model will typically underestimate noise at mm-wave frequencies; however, with only a small modification to include the additional noise sources and account for increased short-channel noise, 60-GHz noise performance can be well-predicted. This amplifier dissipates 36 mA from a 1.5-V supply.

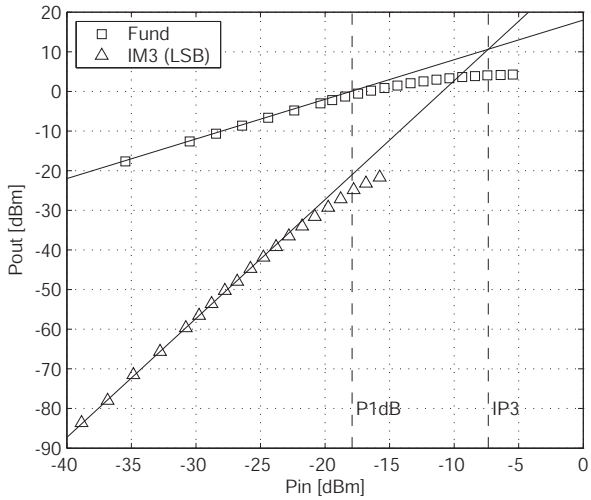


Figure 4.12 Measured two-tone (38 GHz and 38.25 GHz) distortion for the 40-GHz amplifier.

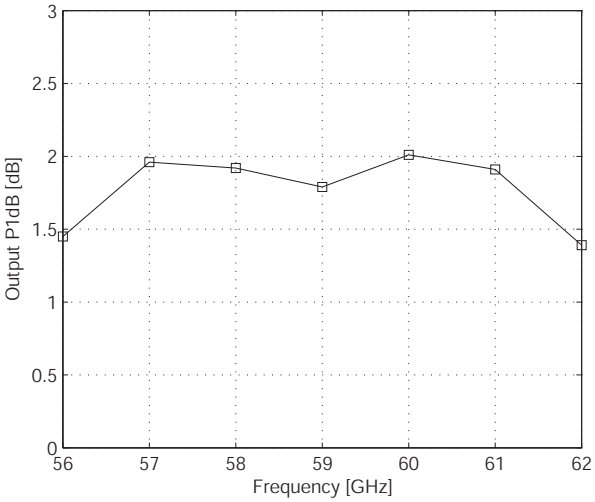


Figure 4.13 Measured output 1-dB compression point for the 60-GHz amplifier.

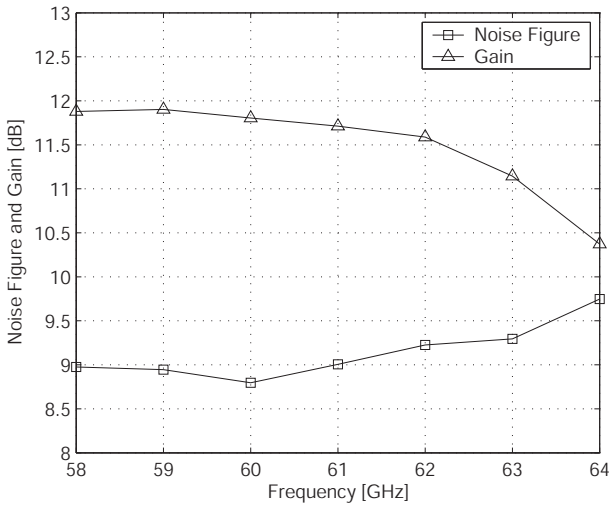


Figure 4.14 Measured noise figure and gain for the 60-GHz amplifier.

Appendix

A4.1 SMALL-SIGNAL MEASUREMENT SETUP

On-wafer *S*-parameter measurements up to 65 GHz were performed using a Cascade Microtech probe station, GSG coplanar probes, and an Anritsu 37397C VNA. Open-short de-embedding [51] was used to remove the effects of the pads when measuring individual devices (transistors, transmission lines, etc.). Pad removal was not necessary for the filter or amplifiers since the pads were incorporated into the design.

The VNA was also used for single-tone compression measurements up to 65 GHz. A 65-GHz Anritsu SC6230 power sensor and Anritsu 2437A power meter was used to calibrate out the losses of the cabling and probes.

A4.2 LARGE-SIGNAL HARMONIC MEASUREMENT SETUP

On-wafer vector-corrected two-tone distortion measurements were performed on a custom setup using the procedure described in [60]. The measurement system was limited to 50 GHz, so the intermodulation distortion was characterized only for the 40-GHz amplifier. A reflectometer external to the VNA was mounted on the probe station to maximize dynamic range. Both on-wafer and coaxial calibration standards are needed to obtain the vector-corrected power at the probe tip. A full description of the procedure and algorithm for de-embedding can be found in [60].

To characterize the nonlinearity at 60-GHz, a different technique was needed. Two common approaches to characterize transistor mm-wave nonlinearity are with mm-wave load-pull measurements and power spectrum analysis [61]. While the former requires automated tuners, the latter, which was chosen in this work, can be performed with only a synthesizer, VNA, and power meter. The fabricated device is driven over a wide range of input power and bias conditions, while the output powers at the fundamental and harmonic frequencies are measured.

In the harmonics power measurement setup, the VNA is configured as a receiver in the Set-On mode. In this mode of operation, the source lock circuitry of the 37397C is completely bypassed which allows all of the 37397C samplers to operate over their full dynamic range. The 65-GHz synthesizer and 37397C are locked to the same 10-MHz reference, enabling coherent reception at the harmonic frequencies. A 65-GHz high dynamic-range power sensor is used to de-embed the insertion loss of the cables, probes, adaptors, etc. from the measurements. Port2 of the VNA requires a 10-dB attenuator to avoid compression when the fundamental power is strong, thus limiting the sensitivity of the power measurement at 60 GHz to approximately –35 dBm.

A4.3 NOISE FIGURE MEASUREMENT SETUP

Noise figure (NF) measurement of the 60-GHz amplifier was performed using a Millitech WR-15 noise source, WR-15 waveguide probes, output isolator, OML 50–75-GHz DSB downconversion mixer, and an Agilent N8973A NF measurement system, as shown in Fig. 4.15.

Figure 4.15 On-wafer noise figure measurement setup.

5

Conclusions and Future Work

5.1 Research Summary

This dissertation describes a mm-wave design and modeling methodology for CMOS devices and circuits. The focus of this work has been on demonstrating the feasibility of 60-GHz circuits using a standard digital CMOS process and developing a deep understanding of the frequency limits of the technology, rather than circuit-level optimizations.

Optimization of the active and passive components for mm-wave circuits leads to transistors with finger widths < 1 μm and the use of CPW transmission lines. There is a tradeoff between device model flexibility and accuracy, and in this work, more emphasis was placed on obtaining accurate and predictable models. The constraint imposed was that all devices were common-source devices of either single NMOS or cascode transistors. While this limits the possible circuit topologies, the result is a high precision model that allows circuits to be designed to operate near the limit of that device.

Given this constraint, a modeling methodology using relatively simple models that carefully account for the resistive losses, has been described and applied to transistors and transmission lines. The models have been verified to achieve broadband accuracy from dc to 65 GHz. Finally, two wideband mm-wave CMOS amplifiers, operating at 40 GHz and 60 GHz, were fabricated using a 130-nm bulk-CMOS technology. The measured *S*-parameters correspond extremely well to the simulated models owing to the accurate device modeling. The 40-GHz amplifier achieves 19-dB gain, output P_{1dB} = −0.9 dBm, and IIP_3 = −7.4 dBm, while consuming 36 mW. The 60-GHz amplifier achieves 12-dB gain, output P_{1dB} = +2.0 dBm, and NF = 8.8 dB, while consuming 54 mW. These were the first published amplifiers operating above 30 GHz fabricated using a mainstream bulk-CMOS technology.

5.2 Future Work

As with any research project, this dissertation is not an end, but represents just a single step in the continually evolving field of integrated circuits. By answering one question, namely that it is possible to design 60-GHz circuits using a bulk-CMOS technology, multiple new questions are generated. Since the original publications [9][13] demonstrated the feasibility of 60-GHz CMOS circuits, there have been many advancements in the field of mm-wave CMOS circuit design, with good progress in some of these areas. Others still remain open areas for investigation.

Scalable mm-wave transistor models There is always a tradeoff between the accuracy of a device model and the flexibility of its use. It is clear that the foundry-provided transistor models are grossly inadequate for 60-GHz circuit design, since many of the key mm-wave parasitics are not included. However, it is certainly conceivable that relaxing the constraint to allow for scalable

transistors and different circuit topologies (e.g., common-gate, differential) will result in a model that is still sufficiently accurate for mm-wave circuit design. It is important to remember that the device model should not be considered in isolation, but that the connections to the device are also critical to ensuring that the model does not change when embedded within a circuit. A further benefit of having a scalable model is that it removes the requirement of being able to accurately de-embed the device measurement. This could enable the use of smaller transistors, which are difficult to measure accurately, for improved power consumption.

CMOS power amplifiers CMOS scaling improves amplifier noise performance and gain at mm-wave frequencies but only exacerbates the difficulty of generating sufficient output power at the transmitter given the lower breakdown voltages. Novel circuit topologies for power combining may be required to generate sufficient output power with low supply voltages. Another approach is to use a spatial power combining scheme to ameliorate the requirements on individual power amplifiers. Compared to corporate power combining structures, which become excessively lossy and bulky as the number of devices increases, the combining efficiency of spatial power combiners is approximately independent of the number of individual power amplifiers [62].

Fully-integrated 60-GHz CMOS transceivers Once all of the individual circuit blocks have been designed, the natural next step is to integrate everything into a 60-GHz single-chip CMOS transceiver. The power consumption and die size of the chip should be optimized while maintaining the necessary gain, noise figure, and linearity for system operation. Other possible areas of research specific to designing a 60-GHz beamforming transceiver include: phase shifter architecture and design (LO, RF, IF),

frequency plan and transceiver architecture (direct conversion, heterodyne, low-IF), and the design of transmit/receive (T/R) switches on CMOS.

Antennas and packaging technology For an antenna array, *N* transceivers will need to be integrated either directly on-chip or into a low-cost mm-wave package. Low-temperature co-fired ceramic (LTCC) substrates offer a promising packaging option due to their low cost and good mm-wave performance. Low-loss transmission lines and efficient antennas operating at mm-wave frequencies have been demonstrated on LTCC by other researchers [63]. Flip-chip bonding of the chip to the LTCC substrate will minimize the parasitic inductances.

There are clearly many more questions that still need to be answered before a 60-GHz CMOS transceiver is realized that meets the consumer electronics requirements for high-speed wireless systems. However, now that a digital CMOS process has been shown to be capable of good performance at 60-GHz, questions that were once thought to be academic are now receiving considerable attention. With the industry backing of standardization efforts, it should only be a matter of time before we see widespread deployment of fully-integrated 60-GHz CMOS transceivers.

References

[1] Proakis, J. G. *Digital Communications, 3rd ed*. Boston: McGraw-Hill, 1995.

[2] R. L. Van Tuyl, "Unlicensed millimeter wave communications. A new opportunity for MMIC technology at 60 GHz," in *IEEE GaAs IC Symp. Dig.*, Nov. 1996, pp. 3–5.

[3] WirelessHD, LLC. [ONLINE]. Available: http://www.wirelesshd.org

[4] IEEE 802.15 Working Group for WPAN [ONLINE]. Available: http://www.ieee802.org/15/

[5] IEEE 802.11 Working Group for WLAN [ONLINE]. Available: http://www.ieee802.org/11/

[6] Wireless Gigabit Alliance [ONLINE]. Available: http://wirelessgigabitalliance.org/

[7] J. M. Gilbert, C. H. Doan, S. Emami, and C. B. Shung, "A 4-Gbps uncompressed wireless HD A/V transceiver chipset," *IEEE Micro*, vol. 28, pp. 56–64, Mar. 2008.

[8] M. R. Williamson, G. E. Athanasiadou, and A. R. Nix, "Investigating the effects of antenna directivity on wireless indoor communication at 60 GHz," in *The 8th IEEE Int. Symp. on PIMRC*, Sept. 1997, pp. 635–639.

[9] C. H. Doan, S. Emami, A. M. Niknejad, and R. W. Brodersen, "Design of CMOS for 60GHz applications," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2004, pp. 440-441.

[10] L. M. Franca-Neto, R. E. Bishop, and B. A. Bloechel, "64GHz and 100GHz VCOs in 90nm CMOS using optimum pumping method," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2004, pp. 444–445.

[11] R.-C. Liu, H.-Y. Chang, C.-H. Wang, and H. Wang, "A 63GHz VCO using a standard 0.25μm CMOS process," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2004, pp. 446–447.

[12] M. Tiebout, H.-D. Wohlmuth, and W. Simbürger, "A 1V 51GHz fully-integrated VCO in 0.12μm CMOS," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2002, pp. 238–239.

[13] C. H. Doan, S. Emami, A. M. Niknejad, and R. W. Brodersen, "Millimeter-wave CMOS design," *IEEE J. Solid-State Circuits*, vol. 40, pp. 144–155, Jan. 2005.

[14] S. Emami, C. H. Doan, A. M. Niknejad, and R. W. Brodersen, "A 60-GHz down-converting CMOS single-gate mixer," in *IEEE RFIC Symp. Dig.*, June 2005, pp. 163–166.

[15] S. Emami, C. H. Doan, A. M. Niknejad, and R. W. Brodersen, "A highly integrated 60GHz CMOS front-end receiver," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2007, pp. 190-191.

[16] I. Bahl and P. Bhartia, *Microwave Solid State Circuit Design, 2nd ed*. Hoboken, NJ: Wiley, 2003.

[17] R. Spence, *Linear Active Networks*. London: Wiley, 1970.

[18] S. J. Mason, "Power gain in feedback amplifiers," *IRE Trans. Circuit Theory*, vol. CT-1, pp. 20–25, June 1954.

[19] C. Enz, "An MOS transistor model for RF IC design valid in all regions of operation," *IEEE Trans. Microwave Theory Tech.*, vol. 50, pp. 342–359, Jan. 2002.

[20] S. Emami, C. H. Doan, A. M. Niknejad, and R. W. Brodersen, "Large-signal millimeter-wave CMOS modeling with BSIM3," in *IEEE RFIC Symp. Dig.*, June 2004, pp. 163–166.

[21] Agilent IC-CAP 2002 User's Guide. [ONLINE]. Available: http://eesof.tm.agilent.com

[22] P. R. Gray and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits, 3rd ed*. New York: Wiley, 1993.

[23] A. van der Ziel, *Noise in Solid State Devices and Circuits*. New York: Wiley, 1986.

[24] D. K. Shaeffer and T. H. Lee, "A 1.5-V, 1.5-GHz CMOS low noise amplifier," *IEEE J. Solid-State Circuits*, vol. 32, pp. 745–759, May 1997; *ibid*. "Correction," vol. 40, pp. 1397–1398, June 2005.

[25] A. J. Scholten, L. F. Tiemeijer, R. van Langevelde, R. J. Havens, A. T. A. Zegers-van Duijnhoven, and V. C. Venezia, "Noise modeling for RF CMOS circuit simulation," *IEEE Trans. Electron Devices*, vol. 50, pp. 618–632, Mar. 2003.

[26] D. G. Yee, "A design methodology for highly-integrated low-power receivers for wireless communications," Ph.D. Dissertation, University of California, Berkeley, Electronics Research Laboratory, Memorandum No. UCB/ERL M01/17, 2001.

[27] H. Statz, H. Haus, and R. A. Pucel, "Noise characteristics of gallium arsenide field-effect transistors," *IEEE Trans. Electron Devices*, vol. ED-21, pp. 549–561, Sept. 1974.

[28] R. A. Pucel, H. A. Haus, and H. Statz, "Signal and noise properties of gallium arsenide microwave field-effect transistors," in *Advances in Electronics and Electron Physics*, vol. 38. New York: Academic Press, 1975, pp. 195–265.

[29] M. W. Pospieszalski, "Modeling of noise parameters of MESFET's and MODFET's and their frequency and temperature dependence," *IEEE Trans. Microwave Theory Tech.*, vol. 37, pp. 1340–1350, Sep. 1989.

[30] J. Stenarson, M. Garcia, I. Angelov, and H. Zirath, "A general parameter-extraction method for transistor noise models," *IEEE Trans. Microwave Theory Tech.*, vol. 47, pp. 2358–2363, Dec. 1999.

[31] J. Wood and D. E. Root, "Bias-dependent linear scalable millimeter-wave FET model," *IEEE Trans. Microwave Theory Tech.*, vol. 48, pp. 2352–2360, Dec. 2000.

[32] BSIM Homepage. [ONLINE]. Available: http://www-device.eecs.berkeley.edu/~bsim3/

[33] R. van Langevelde, L. F. Tiemeijer, R. J. Havens, M. J. Knitel, R. F. M. Roes, P. H. Woerlee, and D. B. M. Klaassen, "RF-distortion in deep-submicron CMOS technologies," in *IEDM Tech. Dig.*, Dec. 2000, pp. 807–810.

[34] K. Koh, H.-M. Park, and S. Hong, "A spline large-signal FET model based on bias-dependent pulsed I-V measurement," *IEEE Trans. Microwave Theory Tech.*, vol. 50, pp. 2598–2603, Nov. 2002.

[35] D. M. Pozar, *Microwave Engineering, 2nd ed*. New York: Wiley, 1998.

[36] C. P. Yue and S. S. Wong, "On-chip spiral inductors with patterned ground shields for Si-based RF IC's," *IEEE J. Solid-State Circuits*, vol. 33, pp. 743–752, May 1998.

[37] F. Ellinger, "26–42 GHz SOI CMOS low noise amplifier," *IEEE J. Solid-State Circuits*, vol. 39, pp. 522–528, Mar. 2004.

[38] T. C. Edwards and M. B. Steer, *Foundations of Interconnect and Microstrip Design, 3rd ed.* New York: Wiley, 2000.

[39] B. Kleveland, C. H. Diaz, D. Vook, L. Madden, T. H. Lee, and S. S. Wong, "Exploiting CMOS reverse interconnect scaling in multigigahertz amplifier and oscillator design," *IEEE J. Solid-State Circuits*, vol. 36, pp. 1480–1488, Oct. 2001.

[40] C. H. Doan, S. Emami, A. M. Niknejad, and R. W. Brodersen, "Millimeter-wave CMOS device modeling and simulation," in *Proc. 2004 IEEE Int. Symp. Circuits and Syst.*, May 2004, pp. 524-527.

[41] G. Carchon, W. De Raedt, and B. Nauwelaers, "Novel approach for a design-oriented measurement-based fully scalable coplanar waveguide transmission line model," *IEE Proc. Microwave, Antennas Propagat.*, vol. 148, pp. 227–232, Aug. 2001.

[42] S. Ramo, J. R. Whinnery, and T. Van Duzer, *Fields and Waves in Communication Electronics, 3rd ed*. New York: Wiley, 1994.

[43] S. A. Wartenberg, "Selected topics in RF coplanar probing," *IEEE Trans. Microwave Theory Tech.*, vol. 51, pp. 1413–1421, Apr. 2003.

[44] J. Browne, "On-wafer probes test Si ICs to 110 GHz," *Microwaves & RF*, vol. 42, pp. 84–90, Apr. 2003.

[45] D. F. Williams, A. C. Byers, V. C. Tyree, D. K. Walker, J. J. Ou, X. Jin, M. Piket-May, and C. Hu, "Contact-pad design for high-frequency

silicon measurements," in *Proc. IEEE Conf. Electrical Performance of Electronic Packaging*, Oct. 2000, pp. 131–134.

[46] G. F. Engen and C. A. Hoer, "Thru-reflect-line: An improved technique for calibrating the dual six-port automatic network analzyer," *IEEE Trans. Microwave Theory Tech.*, vol. MTT-27, pp. 987–993, Dec. 1979.

[47] R. Gillon, W. Van De Sype, D. Vanhoenacker, and L. Martens, "Comparing high-frequency de-embedding strategies: immittance correction and in-situ correction," in *Proc. IEEE Int. Conf. Microelectronic Test Structures*, Mar. 2000, pp. 241–245.

[48] R. B. Marks, "A multiline method of network analyzer calibration," *IEEE Trans. Microwave Theory Tech.*, vol. 39, pp. 1205–1215, July 1991.

[49] D. F. Williams, R. B. Marks, and A. Davidson, "Comparison of on-wafer calibrations," in *38th ARFTG Conf. Dig.*, Dec. 1991, pp. 68–81.

[50] P. J. van 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)," in *Proc. IEEE Bipolar/BiCMOS Circuits and Technology Meeting*, Sept. 1987, pp. 70–73.

[51] 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 *Proc. IEEE Bipolar/BiCMOS Circuits and Technology Meeting*, Sept. 1991, pp. 188–191.

[52] T. E. Kolding, "A four-step method for de-embedding gigahertz on-wafer CMOS measurements," *IEEE Trans. Electron Devices*, vol. 47, pp. 734–740, Apr. 2000.

[53] Q. Liang, J. D. Cressler, G. Niu, Y. Lu, G. Freeman, D. Ahlgren, R. M. Malladi, K. Newton, and D. L. Harame, "A simple 4-port parasitic de-embedding methodology for high-frequency characterization of SiGe HBTs," in *IEEE RFIC Symp. Dig.*, June 2003, pp. 357–360.

[54] G. Gonzalez, *Microwave Transistor Amplifiers: Analysis and Design, 2nd ed*. Upper Saddle River, NJ: Prentice-Hall, 1997.

[55] W. R. Eisenstadt and Y. Eo, "*S*-parameter-based IC interconnect transmission line characterization," *IEEE Trans. Components, Hybrids, Manufac. Technol.*, vol. 15, pp. 483–490, Aug. 1992.

[56] H. Rothe and W. Dahlke, "Theory of noisy fourpoles," *Proc. IRE*, vol. 44, pp. 811–818, June 1956.

[57] H. Hillbrand and P. H. Russer, "An efficient method for computer aided noise analysis of linear amplifier networks," *IEEE Trans. Circuits Syst.*, vol. CAS-23, pp. 235–238, Apr. 1976; *ibid*. "Correction," vol. CAS-23, p. 691, Nov. 1976.

[58] R. Q. Twiss, "Nyquist's and Thevenin's theorems generalized for nonreciprocal linear networks," *J. Appl. Phys.*, vol. 26, pp. 599–602, May 1955.

[59] M. Kantanen, M. Lahdes, T. Vaha-Heikkila, and J. Tuovinen, "A wide-band on-wafer noise parameter measurement system at 50–75 GHz," *IEEE Trans. Microwave Theory Tech.*, vol. 51, pp. 1489–1495, May 2003.

[60] B. Hughes, A. Ferrero, and A. Cognata, "Accurate on-wafer power and harmonic measurements of mm-wave amplifiers and devices," in *IEEE MTT-S Int. Microwave Symp. Dig.*, June 1992, pp. 1019–1022.

[61] I. Angelov, H. Zirath, and N. Rorsman, "Validation of a nonlinear transistor model by power spectrum characteristics of HEMT's and MESFET's," *IEEE Trans. Microwave Theory Tech.*, vol. 43, pp. 1046–1052, May 1995.

[62] R. A. York, "Some considerations for optimal efficiency and low noise in large power combiners," *IEEE Trans. Microwave Theory Tech.*, vol. 49, pp. 1477–1482, Aug. 2001.

[63] A. Yamada, E. Suematsu, K. Sato, M. Yamamoto, and H. Sato, "60GHz ultra compact transmitter/receiver with a low phase noise PLL-oscillator," in *IEEE MTT-S Int. Microwave Symp. Dig.*, June 2003, pp. 2035–2038.