University of California
Santa Barbara

Design and Characterization of Circuits for Next-Generation Wireless Communications Systems

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Electrical & Computer Engineering

by

Robert Maurer

Committee in charge:

Professor Mark Rodwell (Advisor), Chair
Professor James Buckwalter (Committee Member)
Professor Larry Coldren (Committee Member)
Dr. Miguel Urteaga (Committee Member)

June 2017

The Dissertation of Robert Maurer is approved.

Professor James Buckwalter (Committee Member)

Professor Larry Coldren (Committee Member)

Dr. Miguel Urteaga (Committee Member)

Professor Mark Rodwell (Advisor), Committee Chair

June 2017

Design and Characterization of Circuits for Next-Generation Wireless Communications Systems

Robert Maurer

Acknowledgements

I've been incredibly privileged to spend the last 6 years working alongside so many brilliant minds from all over the world in the Rodwell group in beautiful Santa Barbara. Never in my life have I met anyone with the level of dedication and passion for their work that Professor Mark Rodwell has - not only as a scientist and an engineer, but as a teacher and a mentor. To him, I express my deepest and most sincere gratitude for giving me the opportunity to be here and making this all possible. He's relentlessly enthusiastic for his endless quest to learn, innovate, and succeed, but he always prioritizes the well-being of his students first and foremost. I believe this is a truly remarkable quality that deserves acknowledgement. I would also like to thank my PhD committee members, Professor James Buckwalter, Professor Larry Coldren, and Dr. Miguel Urteaga for their valuable advice and criticism during my qualifier and dissertation.

I'm extremely grateful to all of my Rodwell group colleagues, past and present, who have helped me and supported me along the way. Our post-doc, Dr. Seong-Kyun Kim has taught me a tremendous amount about IC design and measurements and he deserves special acknowledgement for the role he played in my academic development. Hyun-Chul Park, Thomas Reed, Zach Griffith, Munkyo Seo, and Colin Sheldon laid much of the groundwork upon which I built my research. The other circuit designers, Arda Simsek, Ahmed Ahmed, Hai Yu, and Ali Farid have all played a significant role in this work as well. I'd also like to thank my colleagues and mentors on the devices team from my past life working in device fabrication - Andy Carter, Johann Rode, Prateek Choudhary, Hanwei Chiang, Cheng-Ying Huang, Sanghoon Lee, Jeremy Law, Evan Lobisser, Vibhor Jain, and Doron Elias. Our newest devices team members, Brian Markman, Yihao Fang, Hsin-Ying Tseng, and Jun Wu have also been a tremendous help.

I'd like to thank the entire cleanroom staff, especially Brian Thibeault, Bill Mitchell,

Tony Bosch, Don Freeborn, Biljana Stamenic, Brian Lingg, Aiden Hopkins, and Tom Reynolds for their tireless work and interest in helping the students learn and succeed.

A special thanks to Teledyne Scientific and Global Foundries for fabricating the ICs used in my research.

Last but not least, I'd like to thank my family - my parents, Dan and Kathy, and my sister, Stephanie, provided me with unconditional love and support through my whole life. My loving wife, Barbara has stood by me, supporting and encouraging me through the most trying times of my career. Without her, I don't believe I'd be here at this point.

Curriculum Vitæ

Robert Maurer

Education

June 2011 - Present M.S/Ph.D. in Electrical and Computer Engineering (Expected), University of California, Santa Barbara

August 2007- May 2011 B.S. in Electrical Engineering, University of Notre Dame

Publications

R Maurer, S-K Kim, M Urteaga, MJW Rodwell, Ultra-wideband mm-Wave InP Power Amplifiers in 130 nm InP HBT Technology, Compound Semiconductor IC Symposium (2016)

R Maurer, S-K Kim, M Urteaga, MJW Rodwell, High-linearity W-band Amplifiers in 130 nm InP HBT Technology, Compound Semiconductor IC Symposium (2016)

S-K Kim, **R Maurer**, A Simsek, M Urteaga, MJW Rodwell, Ultra-Low-Power Components for a 94 GHz Transceiver, Compound Semiconductor IC Symposium (2016)

S-K Kim, **R Maurer**, A Simsek, M Urteaga, MJW Rodwell, A High-Dynamic-Range W-band Frequency-Conversion IC for Microwave Dual-Conversion Receivers, Compound Semiconductor IC Symposium (2016)

Abstract

Design and Characterization of Circuits for Next-Generation Wireless Communications Systems

by

Robert Maurer

Demand for wireless data transfer has been increasing rapidly with the rise of smart devices and mobile video streaming. With dozens of wireless applications currently in use and only a finite bandwidth to work with, engineers are challenged to both expand the upward frequency limit of high-performance, high-efficiency wireless systems and to increase the spectral efficiency of the frequency bands already in use. The development of deep sub-μm silicon-on-insulator transistor technology and powerful computer-aided circuit designing tools have allowed us to create more affordable silion-based phased array ICs at frequencies previously achievable by only military applications. The 5th generation of mobile systems (5G) is now expected to use this type of IC to offer increased wireless data capacity in densely-populated areas using mm-wave frequencies. Demand for wireless data is only expected to continue rising, particularly as new IoT applications such as autonomous vehicles become commercially viable.

The work presented in this dissertation addresses both the need for expanding the usable frequency spectrum and the need to increase spectral efficiency in available bands. It includes a design for an analog beamforming matrix for a spatially multiplexed phased array receiver in silicon SOI technology, low-power high-linearity w-band amplifiers in InP HBT technology, and ultra-wideband mm-wave power amplifiers in InP HBT technology. Spatially multiplexed phased array transceivers have the potential to greatly increase the spectral efficiency of mm-wave frequency bands by re-using frequency spectrum for many

data channels. This type of system can be used to create short-range high-capacity line-of-sight wireless backhaul for crowded city squares or event venues. Mm-wave power amplifiers and high-linearity amplifiers in new 130 nm InP HBT technology represent an IC performance boost which pushes the frequency limits of feasible power-efficient wireless systems.

The measured power amplifier ICs produce output power of larger than 16.5 dBm at the 3-dB gain compression condition from 50 GHz to 100 GHz, and a small signal gain of 15 dB over a 90 GHz 3-dB bandwidth. The peak power-added efficiency (PAE) is larger than 8% over that same frequency range. At 90 GHz, the ICs produce 22 dBm of saturated output power and 14.7% PAE. The measured high-linearity amplifier ICs demonstrate an output-referred 3rd order intercept (OIP3) of 22 dBm, a gain of 6.4 dB, and a noise figure below 7 dB at 100 GHz. New designs for an analog MIMO beamforming matrix IC, a 100-165 GHz power amplifier, and an improved w-band high-linearity amplifier are also outlined in this dissertation.

Contents

Chapter 1

Background

1.1 Introduction

In the 1890s, wireless communication was transforming from a laboratory novelty into a practical method of broadcast and communication [1]. Much earlier in 1865, James Clerk Maxwell predicted the existence of electromagnetic waves through Maxwell's equations. This was experimentally confirmed by Heinrich Hertz in 1888, which showed that wireless communications were possible. By the early 1900s, Guglielmo Marconi was building stations capable of wirelessly transmitting telegraph messages across the Atlantic via electromagnetic waves [1]. Wireless technologies for military radar and communications and commercial broadcasting developed rapidly through the 1900s. The airwaves quickly had to be regulated to avoid interference from multiple signals on the same frequency bands. Demand for wireless at both the consumer level and the military level has only continued to increase as technological developments have opened the door to new applications.

Today, mobile phones are ubiquitous throughout much of the world. In the age of digital information, consumers now want more than the ability to communicate while on

the go. They want access to the internet and video streaming even when they are far away from their WiFi router. The demand for larger overall network capacity continues to skyrocket. The military continues to seek high frequency hardware for applications in detection, imaging, electronic warfare, and battlefield control over communications channels.

There are many architectures of wireless link front-end modules, but almost every modern RF wireless link shares some unifying design principles that are critical for communications engineers and analog circuit designers to understand. In this chapter, I will discuss background information on the basics of integrated circuits (ICs) in modern RF wireless links, network capacity, and circuit techniques, as well as the benchmarks for high-performance designs.

1.2 Basic Architecture of Wireless Links

Figure 1.1 shows a basic block-level illustration of a simple direct-conversion wireless link [1] [2] [3]. The ultimate goal of the system is to take a set of analog or digital data, and send it from one point to another via electromagnetic waves with minimal digital errors or analog distortion. The transmitter contains a mixer, a linear power amplifier, and an antenna. The reciprocal receiver has an antenna, a low-noise amplifier, and a mixer. In this section, I will explain the principles of wireless links which provide a successful transmission of data.

Wireless links are used to transport either an analog set of data in the form of an arbitrary waveform, or digital data in the form of an arbitrary string of bits (1s and 0s) from one place to another. The dataset exists within a complex-valued baseband signal, and must be arranged in a form that can be transmitted within a single channel. Therefore, the data must be encoded, or modulated, into a waveform with a narrow

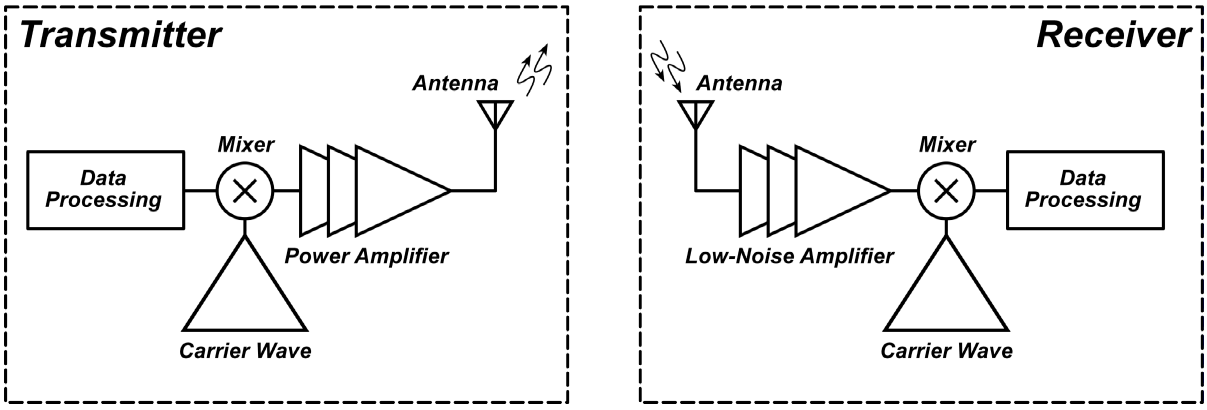

Figure 1.1: Block diagram of a basic wireless link.

channel bandwidth, called a passband [1][4]. The signal is symmetrically de-encoded, or de-modulated, back into the original dataset form at the receive end of the link [4]. There are various types of modulation schemes that are used (AM, FM for analog; PSK, QAM for digital), depending on the application, but all of them use some type of mixer (or multiple mixers) to achieve this goal [1][4]. Mixers and modulation schemes will be discussed further in sections 1.8 and 1.9.

At any point along the wireless link circuitry, the signal carries some average power level. This is expressed mathematically in equation 1.1, where the boldface $\mathbf{V}$ and $\mathbf{I}^*$ are the complex voltage phasor and the complex conjugate of the current phasor respectively [5]. The power of the associated signal can be increased by introducing gain. In modern architectures, gain is achieved via transistor amplification. Depending on the context, gain can refer to an increase in the voltage or current signal amplitude, or the average power level [6]. The design considerations and goals for different types of transistor amplifiers will be discussed in more depth in section 1.7.

$$P_{signal} = \frac{1}{2}\Re(\mathbf{VI}^*) \tag{1.1}$$

A power amplifier (PA) should be used at the output of the transmitter to send the

data set with as much range as possible [1]. The free-space path loss (FSPL) of an electromagnetic wave is defined by equation 1.2, where d is the propagation distance and λ is the wavelength of the carrier wave [7]. This means, for a given channel frequency, the received power drops rapidly with distance, and this problem is exacerbated at higher frequencies. In addition, there is frequency-dependent atmospheric absorption based on the weather and humidity [8]. This is a major design challenge, considering the receiver has a finite sensitivity [1]. In the case of commercial applications, there is an FCC-imposed safety limit for effective isotropic radiated power (EIRP), so power outputs for hardware for these applications are typically designed up to near this limit. The military are not subject to the same set of regulations, and therefore wireless hardware for their applications is often designed for maximum possible transmit power for optimum range and resolution. Power amplifier design considerations and further applications will be discussed more in section 1.4.

$$FSPL = \left(\frac{4\pi d}{\lambda}\right)^2 \tag{1.2}$$

At the receive end, a low-noise amplifier (LNA) is used to maximize receiver sensitivity, thereby maximizing the transmission distance [1]. There is a background level of noise produced by gas molecules in our atmosphere that washes out the desirable signal. Additionally, all transistors, diodes, and resistors generate noise on their own, meaning that at any point along the wireless link, there is a desired signal power and an unwanted noise power [1] [4] [3]. The ratio between these two powers is called the signal to noise ratio, or SNR, which is mathematically expressed in equation 1.3 [5] [1]. As designers, we would like this value to be kept as large as possible to preserve the integrity of the

transmitted data set.

$$SNR = \frac{P_{signal}}{P_{noise}} \tag{1.3}$$

Amplifiers such as PAs and LNAs can be used to increase the power level of the desired signal, however, the noise power will inevitably see the same factor of power increase, or gain [5] [1]. Therefore, since any circuit elements introduce noise into the system, there is a degradation of the signal to noise ratio associated with any circuit block. This degradation is called the noise factor, F, shown in equation 1.4 [5]. This degradation is important, because it sets the minimum acceptable level of received signal power to achieve desired performance at the back end of the receiver [1]. A low-noise amplifier is designed to provide a large amount of gain while simultaneously minimizing the associated degradation of SNR [1]. Design considerations for LNAs will be discussed in greater depth in section 1.5.

$$F = \frac{SNR_{out}}{SNR_{in}} \tag{1.4}$$

Keep in mind that these are just the very most basic building blocks that must be used to achieve reliable wireless transmission. In almost any practical application, there is a local oscillator used to generate the carrier waveform. There are usually multiple filters used throughout the system to block unwanted harmonics and tones from distorting the desired signal. This is also excluding phase-controlling and clock-syncing circuitry used to enable advanced forms of data modulation [1]. These techniques are of critical importance to the function of modern wireless links, however, they will only be peripherally addressed within the scope of this dissertation.

1.3 Linearity and Output Power

One of the important benchmarks for RF performance is linearity. A circuit block is perfectly linear if the transfer function of the amplifier $H(x)$ has the property $H(Ax + By) = AH(x) + BH(y)$ [1]. In qualitative terms, the output signal is a perfect duplication of the input signal, only larger or smaller. In reality, there is always some level of distortion inherent in any amplifier, mixer, or filter which manifests itself in the form of unwanted harmonic tones [1]. Odd-ordered distortion products (3rd order, 5th order, etc), as opposed to even-ordered harmonics, appear at or near the design frequency, and therefore cannot easily be filtered out. As input power is increased, the log of the output power of the 3rd order distortion product signal increases at 3 times the rate as that of the log of the fundamental tone. This is the reason why bias-point amplifiers are biased with a voltage right between the maximum and minimum voltages of the safe linear operating region. It allows the amplifier to handle a larger voltage swing without causing high levels of distortion. Class-A amplifiers are biased such that current is conducting through the entire voltage swing, which creates low levels of distortion at low power with relatively low drain/collector efficiency, while class-B and class-C are biased such that current is conducted for a fraction of the time. Hence class-B and class-C amplifiers have moderate levels of distortion even at low power levels, but higher efficiency than class-A amplifiers [1] [5].

Linearity is quantified in terms of IP3, or 3rd order intercept, in units of watts or dBm [1] [5]. This can be measured using a 2-tone measurement. If there are 2 slightly offset fundamental tones, f_1 and f_2 which are applied to a circuit block simultaneously, 3rd-order intermodulation products will appear at the frequencies $2f_1 - f_2$ and $2f_2 - f_1$ [1]. The 2 fundamental tones are applied at equal power levels below gain compression, and the output power levels are measured along with the power of the 3rd-order intermodula-

tion tones. The output power levels can be plotted vs. input power in terms of dBm [1]. The 3rd-order intermodulation product power levels will follow along a line with a slope of 3, and the fundamental output powers will follow along a line with a slope of 1 [1]. If those lines are extrapolated, there will be a power where the lines intercept. The input power at which this occurs is referred to as the IIP3 or input-referred 3rd-order intercept and the output power at which this occurs is called the output-referred 3rd order intercept or OIP3 [1] [5].

The most high-performance mixers used in RF receivers and transmitters are passive diode mixers [1]. These require a very high-power signal driving the local oscillator port to reduce the distortion introduced by the mixer [1]. Therefore, when designing a power amplifier to drive the local oscillator port of a passive mixer, output power is sometimes quantified in terms of saturated output power. This is the value of P_{out} at which it no longer increases with an increase in P_{in}.

The boundaries on linearity in a class A amplifier are determined by the breakdown voltage, V_{br}, current cutoff, where I_c or I_d reaches zero, and the collector or drain saturation voltage or knee voltage. As the output signal becomes more distorted, the apparent gain drops, resulting in a saturation of the P_{in} vs. P_{out} curve [5]. Therefore, in the case of a class-A amplifier, the optimum output impedance for high linearity is also the optimum output impedance for maximum output power for a given bias point. The compression in the gain curve is quantified in different ways depending on the purpose which the amplifier serve. The 1-dB compression point, P_{1dB}, is the output power at which the amplifier's gain has reduced by 1-dB from the small signal value, and the 3-dB compression point, P_{3dB} is the output power at which the amplifier's gain has dropped by 3-dB from the small signal value. For driver amplifiers, it is often useful to quantify the saturated output power, P_{sat}, which is the very maximum amount of power the amplifier is capable of emitting. As a transmitting amplifier, the output should have low levels of

distortion to keep the out-of-band interference small, and, in systems with very complex modulation constellations, to avoid bit errors. Therefore, it is common for transmitter amplifiers to quantify output power in terms of 1-dB compression point, P_{1dB} [1]. P_{1dB} is the value of output power at which the power gain is 1 dB lower than it is at small signal [1] [5].

The linearity of a receiver is particularly important for both military communications systems and consumer and commercial applications. In military applications, enemies may be trying to block communications with high-power in-band jammers. High-power signals that fall within the receiver passband are capable of distorting the desired signal and creating unwanted products which also fall within the receiver passband. Therefore, in any battlefield-applicable wireless application, it is desirable to achieve extremely high linearity. In commercial applications, depending on the quality of filtering in the receiver, linearity is very important to prevent interference from nearby frequency bands. Thus, for any application, linearity is a desirable trait for both transmit and receive functions. Oftentimes, the limiting factor on wireless transceiver linearity is the cost of the hardware.

1.4 Noise

If you have ever been in your car searching for a radio station and landed on a channel which is not currently broadcasting, youve heard the loud irritating sound we usually refer to as white noise or radio static. It's reasonable to wonder why you hear this rather than silence. If nothing is broadcast, why does our radio make any sound at all? In addition to the many other channels which are being broadcast simultaneously at other frequencies, there are also thermally-induced electromagnetic fluctuations in the circuit elements of the receiver which produce random voltage and current variations [1] [4] [5]. Furthermore, there is a background level of thermal noise produced in the atmosphere [1]. We refer to

all of these random electromagnetic fluctuations collectively as noise. In general, if we are to successfully transmit a set of data wirelessly, the data signal at the de-modulated end of the receiver must have more power than the noise at that same point, therefore noise determines the minimum power level required for successful wireless data reception [1]. Ideally, the signal power is significantly larger than the noise power so that the probability of errors in the data is minimized. In this section, I will discuss the basics of noise and the role it plays in wireless communications.

The Boltzmann constant is a physical constant which relates the average kinetic energy in particles of gas with the temperature of the gas. The result is that there is a spectral density of noise power produced in the background depending on the background temperature. The background noise power is $kTBW$ where k is the Boltzmann constant, T is the temperature in Kelvin, and BW is the receiver bandwidth. At room temperature, this is $4.14 * 10^{-21}$ Joules per Hertz, or in logrithmic units, $-173.9\frac{dBm}{Hz}$ [1]. The receiver will detect the desired signal with a particular power level depending on the transmit power and distance, antenna directivities, attenuation due to weather, etc, which provides the receiver with an initial SNR. Unfortunately, since any circuit elements introduce additional noise into the system and any amplification of the desired signal will also apply to the noise signal, the SNR cannot be improved and will only ever be degraded [1] [5]. The degradation of the SNR associated with a circuit block is called the Noise Factor F as discussed earlier. This is often expressed in terms of Noise Figure, NF, expressed in dB instead of magnitude, shown in equation 1.5 [5].

$$NF = 10 \log (F) \tag{1.5}$$

The noise signal at the input of a group of series-combined amplifiers is going to be amplified more than noise introduced after gain has been introduced. Equation 1.6 shows

the overall noise factor of N series-combined amplifier stages, where G_m is the magnitude of the gain of stage m [1] [4] [5].

$$F_{total} = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} + ... + \frac{F_N - 1}{G_1 G_2 ... G_{N-1}} \tag{1.6}$$

So the noise contribution from each gain stage is smaller than the stages that came before it. This is why, for the very first gain stage of a receiver, it is critical to use a low-noise amplifier which adds gain while introducing as little noise as possible into the system to maximize sensitivity.

As mentioned earlier, a larger SNR produces a lower probability of errors for each bit transmitted. The bit error rate or BER is the probability of a transmission error for each bit which is transmitted - depending on the application, the acceptable level of BER can be on the order of 10^{-3} or can be even lower than 10^{-12}. When quantifying the sensitivity of a receiver in terms of the lowest acceptable received signal power, we need to first determine the minimum acceptable BER and the minimum acceptable SNR, SNR_{min}.

The sensitivity of a receiver, S_{dBm} [9] [10], in terms of minimum acceptable received signal power is therefore:

$$S_{dBm} = 10 \log_{10} \left(\frac{kTBW}{1mW} \right) + NF_{Receiver} + SNR_{min,dB} \tag{1.7}$$

The dynamic range of a system is the difference between the highest amount of power a system can receive, and the lowest amount of power a system can receive [1]. Therefore, the boundaries of dynamic range are set by the noise floor on the low end and linearity on the high end. This is an important figure of merit for high-performance receivers [1].

1.5 Capacity and Bandwidth

Commercial wireless technology is driven by the demand for increased capacity. At the individual consumer level, people want to talk on their phones, send text messages, surf the internet, and stream videos all from their phone even when they are far away from their WiFi router. Streaming videos wirelessly over mobile channels in particular requires a large data capacity on individual channels. At the network level, mobile phone usage continues to become more ubiquitous. Allowing many people in a local area to simultaneously consume and send data on their mobile phones requires the use of many channels. The FCC must allocate a finite frequency bandwidth for mobile networks so that it does not interfere with other bands reserved for radio, TV, navigation systems, radio astronomy, military, and many other applications. Unfortunately, within a limited bandwidth, it is impossible to limitlessly increase both the number of channels in a given area and the capacity of each channel [1] [11]. The limits of bandwidth and capacity and the ways to improve them will be discussed in this section.

The foundation of communication theory is the Shannon Capacity, C. For a channel with additive white Gaussian noise, of bandwidth BW and signal to noise ratio of SNR, the capacity is in equation 1.8 [11].

$$C = BW log_2 (1 + SNR) \tag{1.8}$$

In 1948, Claude Shannon showed that the fundamental limit of bits per second for a given band-limited noisy channel is directly related to the bandwidth and the signal-to-noise ratio of the signal [11]. Since the bandwidth of a network is typically fixed by standards and regulations, this makes it the designers job to design links with the largest possible SNR within a given power budget, maximizing spectral efficiency (data rate per bandwidth) [1]. However, in the context of next-generation technologies, it is important

to note that developing viable circuit and device technologies at previously unreachable frequencies gives the FCC more bandwidth to allocate in the future, which will ultimately play an important role in the expansion of network capacities. This is a concept that will be addressed in this dissertation, particularly in the 3rd chapter.

It is important to note that Shannon's proof relies on channel codes of arbitrarily great length, and does not show how to attain this capacity with practical codes of moderate length and hence moderate computational complexity within the receiver. Within the scope of this thesis, I will not devote much attention to the various ways that modulation can be used to minimize the bit error rate at Shannon capacity. Instead, more attention will be devoted to ways the limit can be increased through bandwidth expansion and spatial multiplexing techniques which can efficiently improve spectral efficiency.

Historically, channels are divided up using frequency multiplexing (figure 1.2) within the allotted bandwidth for a local mobile network, although Code Division Multiple Access (CDMA) uses a set of non-orthogonal codes [1]. For hardware designers, the total allowable frequency band has already been pre-determined by regulations, so to accommodate a large number of users, many channels with narrow bandwidths are employed within the range of each cell tower. In urban areas with high population densities, this becomes a major issue as mobile data usage becomes more ubiquitous. One possible solution to increase network capacity is to increase the number of channels within a given bandwidth by using spatial multiplexing [12]. This can be done using phased antenna arrays, which will be discussed in more depth in the next chapter.

Expanding usable frequency limits to open up more usable bandwidth is limited at the circuit level by limited device performance at high frequencies [1] [5]. In the consumer market, device limitations at high frequencies are especially problematic. Since commercial RF ICs are designed with the intention of feasible mass production, they are often designed in silicon CMOS technology, which is cheap and can be produced entirely

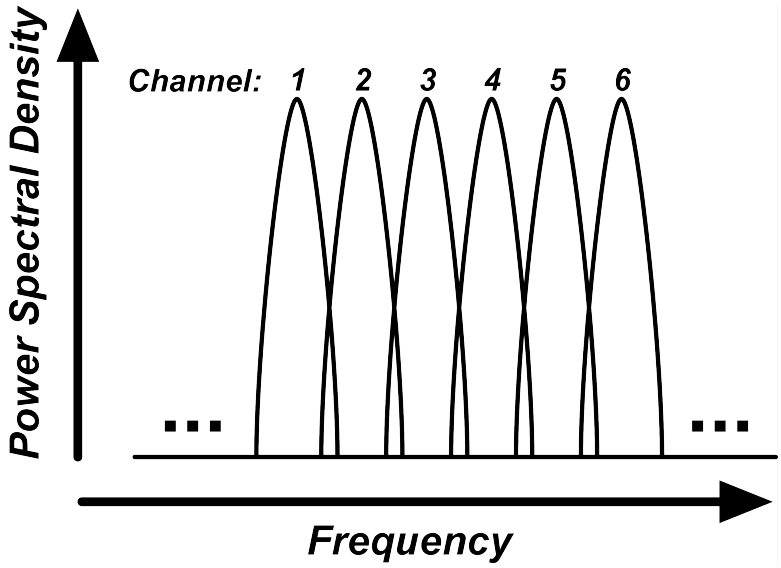

Figure 1.2: Frequency multiplexing involves allocating discrete sections of network bandwidth to separate different channels

with optical lithography. RF ICs for military applications and instrumentation are often developed less cost-effective process technologies, such as InP HBT technology, which is much faster, but is capable of high performance at much higher frequencies [13].

Another method opening up wireless communication channels at higher frequencies is with the use of highly directional phased antenna arrays to combat severe free space path losses and atmospheric absorption at higher frequencies [14]. Any antenna has a radiation pattern which determines how much of the delivered power is radiated in any given direction [1]. Only the portion of the beam directly pointing at the receiver is useful, and larger beam widths and secondary lobes of the radiation pattern limit the power of the received signal. Improving the directionality, D, of an antenna on either the transmitter or receiver improves the antenna gain, G, as seen in equation 1.9 [7]. The effect of this on received power can be seen in the Friis Transmission Equation (equation 1.10) [7].

$$G = E_{ant}D \tag{1.9}$$

$$\frac{P_r}{P_t} = G_t G_r \left(\frac{\lambda}{4\pi R}\right)^2 \tag{1.10}$$

where E_{ant} is the antenna efficiency, G_t and G_r are the gain of the transmit and receive antennas respectively, P_t and P_r are the transmit and receive powers, and λ is the wavelength of the signal [7]. As frequency increases, λ decreases, which reduces the received power, limiting the range of the link. The transmit power is typically limited by FCC regulations for many applications, so the most sensible knobs to turn in this equation are the antenna gains. The FCC also limits the effective isotropic radiated power, EIRP, which is equal to $G_t * P_t$. Consequently, at some point, the gain of the transmit antenna will reach a maximum allowed value, and the only knob to turn is the gain of the receive antenna G_r. Increasing the directionality of the antennas can improve SNR and enable channels at higher frequencies, which will improve network capacity [11] [7]. This is not a simple task, as most antennas are mechanically stationary, and a directional antenna needs to be pointed at its target. This is difficult for a mobile phone user who is moving around or does not have a direct line of sight to the tower. Directional antennas for mobile usage will be addressed in the next chapter.

1.6 Microwave and Mm-wave Design

Extending usable bandwidth to higher frequencies requires advanced circuit design techniques beyond traditional analog design methods [1] [5]. As wavelengths shrink to dimensions comparable to wire lengths, the inductive and capacitive interconnect parasitics must be carefully tracked and taken into design consideration to prevent unwanted reflections and impedance transformations [5]. At these frequencies, because interconnect inductive and capacitive parasitics are significant, interconnects with controlled

impedances are used to propagate confined electromagnetic modes. This way, the parasitic effects are predictable and can be included in the IC design. As frequencies continue to increase, even smaller, previously insignificant parasitic capacitances and inductances also begin to have a larger impact on circuit and device performance. In the past few decades, advances in computer aided design, electromagnetic modeling software, fabrication technology, and computational power have enabled reliable integrated circuit designs at frequencies larger than 100 GHz. As computing power and RF device frequency performance continue to progress, there is a path forward to improving the available bandwidth of wireless technologies. Circuit designers will be required to use techniques described in this section to make more energy-efficient designs

Typically, in dealing with circuit blocks at high frequencies, scattering parameters, or S-parameters, are used most often, and are often supplemented with Y- and Z- parameters for device model extraction [5]. Figure 1.3 shows how scattering parameters are quantified. Here, we are quantifying the amplitude of incident waves at port n with the variable a_n and the amplitude of the outgoing wave at port n with the variable b_n [5]. The S-parameters for a 2-port network can be represented by the matrix shown in equation 1.11 [5]. Since S_{21} is the amplitude of the wave outgoing from port 2 of a 2-port network divided by the amplitude of the incident wave to port 1, it follows that S_{21} is the insertion voltage gain (or loss) of a 2-port network and $|S_{21}|^2$ is the insertion power gain [5].

$$\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} \frac{b_1}{a_1} & \frac{b_1}{a_2} \\ \frac{b_2}{a_1} & \frac{b_2}{a_2} \end{bmatrix} \tag{1.11}$$

Designing circuits at microwave frequencies requires a solid understanding of analog circuit design principles and impedance tuning. For any circuit, there are optimum impedance conditions for various intended outcomes [1] [5]. For example, to design a

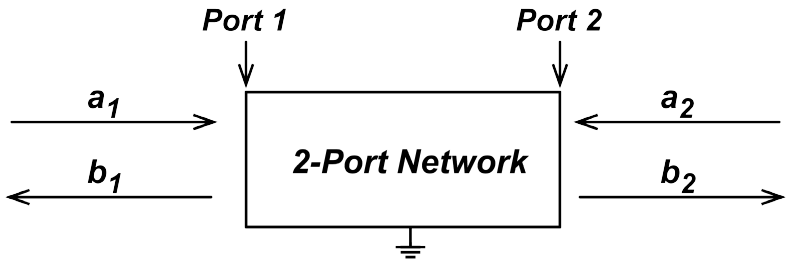

Figure 1.3: 2-port network showing incident and reflected components

two-stage amplifier with maximum power transfer between stages, the designer would like to ensure that the input impedance of the second stage is the complex conjugate of the output impedance of the first stage [6]. In a low-frequency analog design, one generally focuses on voltage gain at a given specified load impedance. In a microwave frequency design, where transistor available power gains can be low, to obtain adequate gain, one may need to provide impedance matching so as to obtain gain close to the maximum available gain. It is efficient to introduce a carefully modeled passive impedance transformation network [1]. To understand why this works, consider equation 1.12 representing the input impedance of a transmission line terminated with a load resistor. β is the phase constant, Z_o is the characteristic impedance of the transmission line, and Z_L is the load resistance [5]. Note that at frequencies where $l << \lambda$, the imaginary component approaches 0, and the input impedance approaches Z_L [5]. As frequency increases and wavelength decreases, it takes less space to provide passive transmission line impedance transformations.

$$Z_{in}(l) = Z_o \frac{Z_L + jZ_o \tan(2\pi l/\lambda)}{Z_o + jZ_L \tan(2\pi l/\lambda)} \tag{1.12}$$

As one could imagine, dealing with these complex expressions for multi-stage circuits and taking minute parasitic elements into account using analytical numeric expressions is extremely unwieldy for designs of any appreciable complexity. This is why Smith

charts and computer-aided electromagnetic modeling are used to quickly visualize and determine impedance or admittance transformations, and take small parasitic values into account. The Smith chart is a graphical representation of normalized complex numbers which can easily transform numbers between impedance-space and admittance-space.

The center of the chart, seen in figure 1.4, is the normalized impedance of 1 + j0. Typically, the system impedance (commonly 50 Ω) is the normalization impedance [5]. The horizontal line across the center of the chart represents a geometric progression of the real component of the impedance where the left corner is a short circuit (no resistance or infinite conductance) and the right corner is an open circuit (infinite resistance or no conductance). Radial lines extend from the right or left side of the chart, representing the lines of constant reactance or susceptance, respectively (figure 1.5a, b). These radial lines of constant reactance or susceptance are always perpendicular to circles of constant resistance or conductance respectively. If S-parameters are plotted on the chart, a parameter or reflection with a magnitude larger than 1 will fall outside the boundaries of the unit circle [5]. In bilateral amplifiers, for example, S_{11} or S_{22} landing outside of the Smith chart is an indication of negative input or output impedance and hence instability [1] [5].

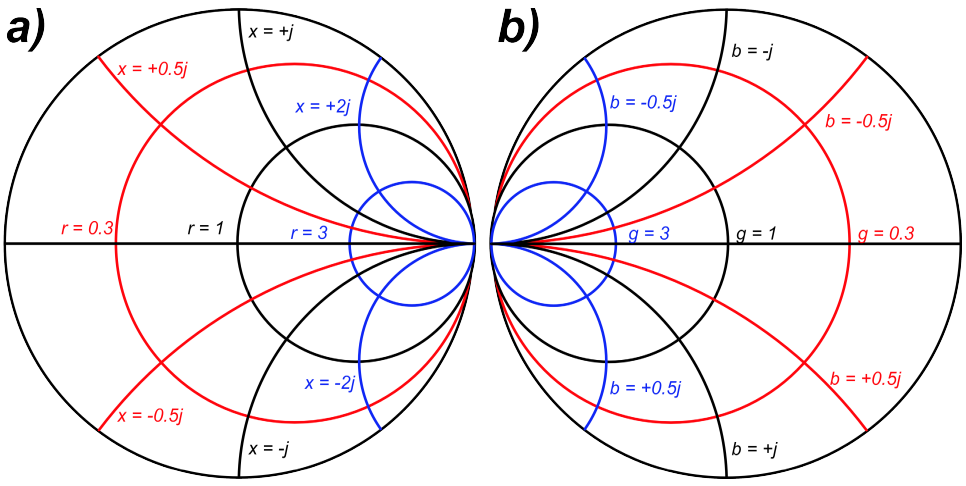

Figure 1.4: Normalized Smith charts in a) Impedance space, b) Admittance space

Stability can also be determined by calculating the Rollet stability factor, K, and the μ stability factor from S-parameters at each frequency as shown in equations 1.13 and 1.14 [5].

$$K = \frac{1 - |S_{11}|^2 - |S_{22}|^2 + |\Delta|^2}{2|S_{21}S_{12}|} \tag{1.13}$$

$$\mu = \frac{1 - |S_{11}|^2}{|S_{22} - S_{11}^*\Delta| + |S_{21}S_{12}|} \tag{1.14}$$

where $\Delta = S_{11}S_{22} - S_{12}S_{21}$ is the radius of the stability circle at that frequency. The circuit is unconditionally stable if $\mu > 1$ and $K > 1$.

When dealing with lumped reactive elements, such as inductors and capacitors, the Smith chart is a helpful tool for visualizing how the impedance is transformed with each element, starting from the load and moving backwards towards the input. For an added inductance, the impedance moves upwards along a circular path by the normalized imaginary impedance associated with the inductor. If the inductor is a shunt element, it moves along the circle of constant conductance, and if it is a series element, it moves along the circle of constant resistance in impedance space. Similarly, adding a shunt or series capacitance results in a downwards move along the circle of constant conductance or constant resistance respectively [5].

Transmission line impedance transformations can also be easily made using Smith charts. A series transmission line with length l creates a clockwise rotation of $\frac{\pi\lambda}{4l}$ radians about the circle centered at the point on the Smith chart representing the normalized characteristic impedance of the transmission line. Adding a shunt transmission line moves the impedance upwards along the circle of constant conductance an angle equal to $\frac{\pi\lambda}{4l}$ radians. Being able to easily plot and visualize impedances this way also makes it easier to design for optimum noise or power transfer conditions. As I will explain in more detail

later, the impedance optimized for different conditions can also be plotted on the Smith chart to help design for high linearity, high output power, or low noise [1] [5].

As design frequencies become a substantial fraction of the transistor f_{max} (maximum frequency of unity power gain), the maximum available stable gain drops [1] [6] [5]. When the transistor gain is high, it allows the designer to use techniques to trade gain for linearity or reduced noise. This makes it more difficult to achieve high linearity, output power, or efficiency while still achieving enough gain to be of use to the overall system and low noise figure. At high frequencies, transistor cell size and bias conditions should be planned carefully to minimize the insertion losses and parasitic elements introduced by passive impedance matching networks. Transmission lines are not entirely lossless, and the associated insertion losses can introduce thermal noise on the input of an LNA or eat into power gains (and therefore power-added efficiency) at the output of a PA. As we continue to push the limits of high frequency circuit design to open up usable wireless bandwidth, these techniques and ideas will continue to play a more important role for circuit designers.

1.7 Transistor Amplifiers

Quantitatively, the vast majority of transistors in use today are used for digital logic, however, for front-end wireless link design, we are generally more interested in the minority which are used for analog amplification of arbitrary waveforms. There are 2 broad categories of transistors - MOSFETs and BJTs, and both of them can be used for either digital or analog applications [6]. In digital logic, circuit blocks are designed to provide either a high (1) or low (0) output, depending on the inputs and functionality of the circuit block. In analog applications, circuit blocks are designed to take an arbitrary waveform and reproduce it with a desired effect. Filters remove unwanted frequencies,

mixers modulate or de-modulate frequencies, and amplifiers prodive linear gain to a signal in a particular frequency range. In this section, I will discuss the basic techniques and concepts used to design amplifiers using transistors.

Gain is the quality of a transistor that makes amplification of an arbitrary waveform possible. This means that a signal applied to one terminal of the device, is reproduced at another terminal, only larger. Gain can refer to voltage gain, current gain, transconductance (current from voltage), transimpedance (voltage from current), or power gain in terms of which quantity of the signal is being amplified. Power gain is simply output power P_{out} divided by input power P_{in}. Gain can vary over a large range of magnitudes, so it is often expressed in logarithmic form in terms of dB, such as for power gain as expressed in equation 1.15 where G_{power} is the amplifier power gain [5].

$$G_{power} = 20 \log \left(\frac{P_{out}}{P_{in}} \right) \tag{1.15}$$

BJTs have a number of advantages over MOSFETs in terms of RF circuit performance. They generally have a higher switching speed for comprably advanced technology nodes, making them preferable for very high-frequency applications [13]. They demonstrate higher drive current per unit of area which makes them superior for power applications. They have a higher output impedance, which makes them better current sources, and they have better noise performance. For applications where mass-production is not an issue, such as high-performance measurement instruments and specialized military-grade equipment, BJTs are used more often. For most consumer applications such as mobile phones, MOSFETs are predominantly used because of cost. The fastest silicon-based CMOS technology can be patterned entirely via optical lithography, making the effective cost per chip very small aside from the design and mask development costs. This brings some extra limitations for circuit designers working in these areas.

The three common forms of amplifier are common emitter/source, common base/gate, and emitter/source follower [6]. These are shown in figure 1.5.

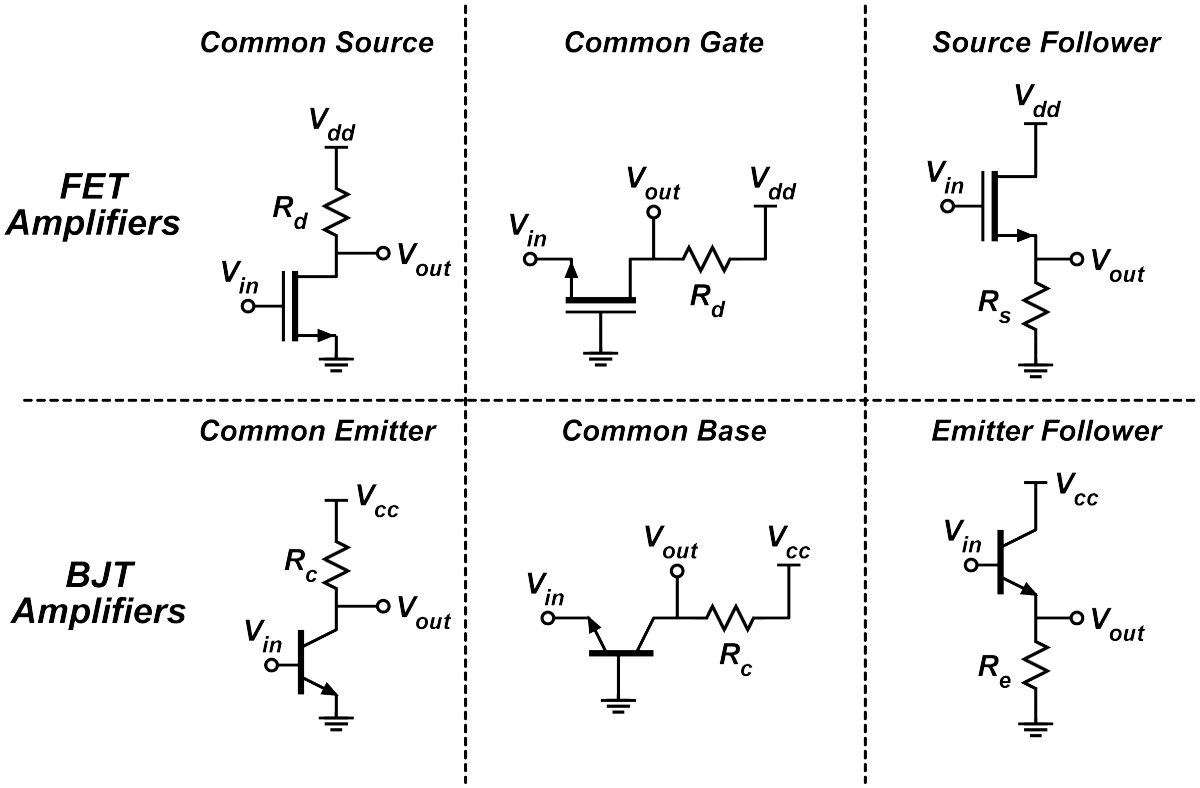

Figure 1.5: Amplifier topologies

Common emitter/source amplifiers are used most often, as they provide a high voltage gain and current gain and a moderate input and output impedance. Common base/gate amplifiers can have a very large voltage gain, but unity or smaller current gain. They have a low input impedance and a high output impedance. Emitter/source follower amplifiers have a voltage gain at or smaller than unity and can have a high current gain. They have a large input impedance and a small output impedance. Common base and common collector configurations are often used as buffer amplifiers or impedance transformers. Each of these amplifier topologies requires that the transistor be biased in forward active mode to function correctly [6].

Transistor amplifiers are almost always designed to provide some type of gain or an impedance transformation, however, depending on the application, they may be designed to optimize for other properties, such as high power or linearity or low noise.

1.7.1 Power Amplifiers

As their name suggests, the key function of a power amplifier is to deliver a large amount of power to a load. Like other transistor amplifiers, it consumes DC power and converts it into AC power under the control of the input signal. The key classes of interest in the scope of this dissertation are class A, class B, and class AB. There are many other classes of power amplifiers, but most of these rely on harmonic cancellation to reduce distortion or increase efficiency, which is not practical at very high frequencies. A class-A amplifier has the lowest levels of distortion, but does not provide gain as efficiently as a class-B amplifier. An example of a class-A amplifier circuit with a bipolar transistor may look similar to a single-stage common emitter or common base amplifier as shown in figure 1.5. Figure 1.6 shows the IV curves in the safe linear operating region for a common-emitter amplifier. The red dashed line represents the current and voltage conditions that should be met through a full 360 degree cycle to achieve the highest possible output power. For class A operation, we assume that the DC voltage and DC current are chosen to such that the quiescent point in figure 1.6 is at the center of the red dashed line. In this case, $V_{dc} = \frac{V_{cc}-V_{min}}{2}$, $I_{dc} = \frac{V_dc}{R_L}$, and $P_{l,dc} = V_{dc}I_{dc} = \frac{(V_{cc}-V_{min})^2}{4R}$ where $P_{l,dc}$ is the DC power consumed by the load. This same amount of power is also consumed by the transistor, since any current flowing through the resistive load also flows through the collector of the transistor and it sees the same voltage swings from the collector to the emitter, so $P_{t,dc} = \frac{(V_{cc}-V_{min})^2}{4R}$. The RMS power supplied to the resistive load is $P_{ac} = \frac{1}{8}V_{pp}I_{pp} = \frac{(V_{cc}-V_{min})^2}{8R}$. Therefore in this ideal condition, the efficiency of

the amplifier is $\eta = \frac{P_{ac}}{P_{l,dc}+P_{t,dc}} = 25\%$. This is the maximum efficiency limit of a class-A amplifier with a purely resistive load with DC coupling between the amplifier and load [6] [1].

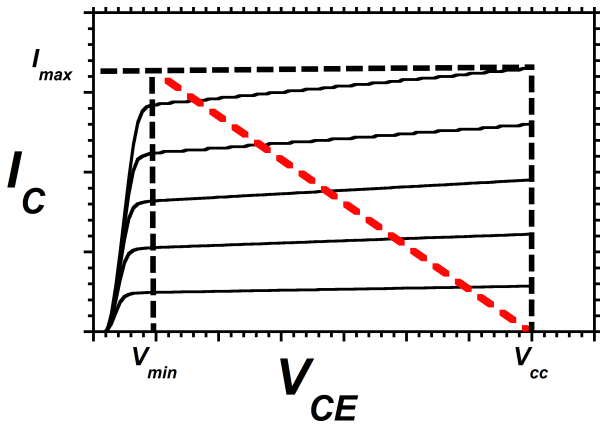

Figure 1.6: Load line for a class-A common emitter amplifier with a purely resistive load

This efficiency value can be improved by AC coupling to the load using a DC blocking capacitor and an inductor as shown in figure 1.7 [1]. This improves the maximum possible efficiency by allowing the DC collector voltage to simply be V_{cc} and by allowing the maximum AC voltage value to be larger than the supply voltage. In this way, the AC power delivered to the load is effectively doubled and the DC power consumption stays the same. This increases the maximum efficiency to 50%. The 50% limit can also be applied if there is an impedance transformation between the collector and the load resistance.

1.8 Mixers

Mixers are frequency conversion elements, and are used in transmitters and receivers to shift the signal information, lying at baseband frequencies, to frequencies surrounding that of the RF carrier [1] [4]. Mixers can be active or passive. A mixer has 3 ports - an RF port, an IF port, and an LO port. In this section I will give a mathematical description of

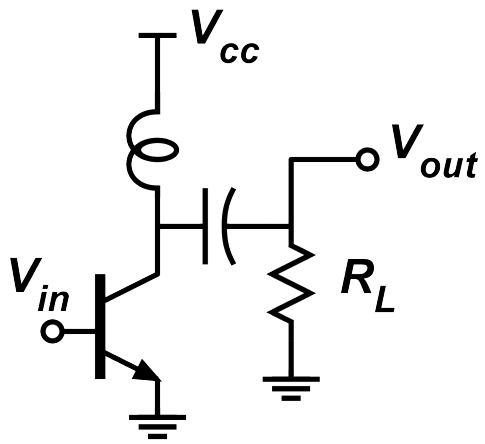

Figure 1.7: Class-A common emitter amplifier with capacitively coupled load for improved efficiency

mixing in the time domain and the frequency domain and discuss the relative advantages and disadvantages of passive diode mixers and active transistor mixers.

A mixer is a non-linear device that multiplies two signals to produce desired harmonics of the two input signals. To demonstrate this mathematically, I will consider the time-domain case of a received RF signal $V_{RF}(t) = S(t) \times A \times \cos(\omega_c t)$ mixing with an LO signal $V_{LO}(t)(t) = \cos(\omega_c t)$ to produce an IF signal which consists of transmitted data lying at the baseband frequency $V_{IF}(t) = S(t)$. We will consider an idealized case where the mixer will simply multiply the RF signal by the LO signal in the time domain. The unwanted harmonics will then be filtered out to isolate the desired signal. A useful trigonometric identity to analyze this is $\cos\theta \times \cos\phi = \frac{1}{2}\cos(\theta - \phi) + \frac{1}{2}\cos(\theta + \phi)$. Note that this results in two terms, each with an amplitude of 1/2, as the amplitude of the input terms are divided among the two output terms. If we use this to multiply $V_{RF}(t) \times V_{LO}(t)$, we get the expression $\frac{S(t)}{2}(1 + \cos(2\omega_c))$. At this point, to isolate the baseband-frequency component, a low-pass filter would be used to remove the unwanted harmonic which lies at twice the carrier frequency. If a phase shift were applied, the amplitude of the

baseband-frequency signal would be multiplied by the cosine of the phase offset. If the LO signal were offset 90 degrees from the RF carrier frequency, the resulting baseband component would be 0.

In the frequency domain, an idealized mixer that sees two input signals at frequencies f_1 and f_2 produces an output signal with components at $|f_1 + f_2|$ and $|f_1 - f_2|$, as multiplication in the time domain is the same as convolution in the frequency domain. In any practical circuit, it is impossible to create this perfect idealized mixer. If we use a switching device like a diode or a transistor, the transfer function does not behave as a perfect multiplier - it also has higher-order nonlinearities in the transfer function that will create higher-order harmonics in the output signal. In the frequency domain, an unbalanced diode or transistor mixer will produce harmonics at any frequency satisfying equation 1.16, where for an Lth-order nonlinearity, $+/-m+/-n = L$ [1].

$$f_{spur} = mf_{RF} + nf_{LO} \tag{1.16}$$

The amplitudes of these tones depend on the amplitudes of the input tones and the coefficients of the higher-order polynomial terms in the transfer function of the mixer. A balanced mixer cofiguration, such as the ring diode mixer in figure 1.8 can be used to eliminate the even-ordered harmonics, although the unwanted odd-ordered harmonics must still be filtered out to isolate the desired output. Passive mixers like this that use diodes as the non-linear mixing element always produce an output signal that has a lower power than the input signals.

An example of an active mixer is shown in figure 1.8. This gilbert cell mixer uses two cross-coupled differential amplifier stages. The differential RF ports drive a differential common-emitter amplifier with RF transistors Q1 and Q2. The differential LO port provides either a 0 or 180 degree phase shift depending on the sign of the LO signal at

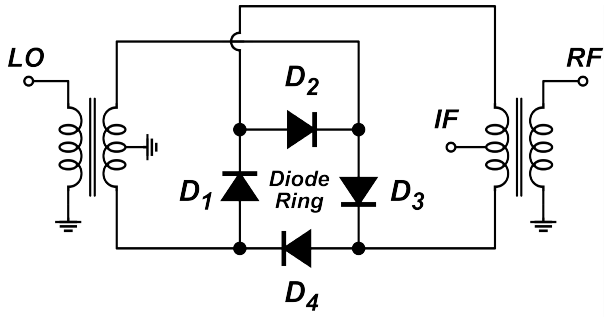

Figure 1.8: Diode ring mixer circuit diagram

a given point in time. An active mixer like this, in contrast to a passive diode mixer, consumes DC power. It also will likely have a much higher noise figure and be more easily overloaded. The key advantage of an active mixer is that it can provide conversion gain, where a passive mixer will always have loss.

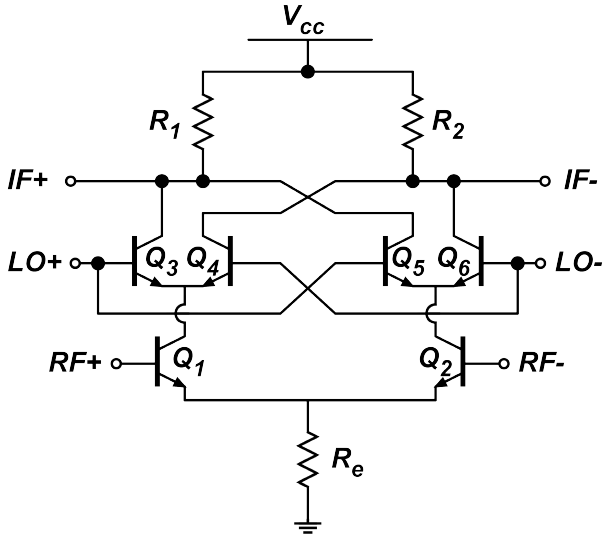

Figure 1.9: Gilbert cell circuit diagram

1.9 Modulation

Wirelessly transmitted data must have a radio carrier frequency (RF) and a separable set of modulated data. Whether arranged in digital or analog form, raw data is

at frequencies much lower than typical RF carrier frequencies, and therefore cannot be transmitted on its own along an individual frequency channel. It is therefore mixed with RF carrier [4]. Common historical modulation formats include AM (Amplitude Modulation) and FM (Frequency Modulation), both widely used in analog radio transmission. Digital schemes include PSK (phase shift keying) and QAM (Quadrature amplitude modulation) [1] [4].

To get a basic understanding of how data can be modulated into an RF signal, let us first briefly look at AM and FM schemes. The goal in this case, is to send auditory data in the form of a bandlimited set of time-dependent sound waves, $s(t)$, within an individual carrier frequency defined by the sinusoidal equation $c(t) = A\sin(\omega_c t)$ shown in figure 1.10. In the case of an AM scheme, the sound wave $s(t)$ will be encoded in the amplitude of the sinusoidal carrier frequency such that the total transmitted wave is $W_{AM}(t) = [A + s(t)]\sin(\omega_c t)$. An FM signal, on the other hand, has a constant amplitude, but has a variable frequency shift determined by the sound wave and a modulation scaling constant that has units of radian per volt. The resulting wave in this case would be $W_{FM}(t) = A\sin[\omega_c t + Ks(t)]$. These waveforms can be seen in figure 1.10. The baseband signal $s(t)$ is called the complex baseband representation or complex envelope [4].

Digital data is generally transmitted as PSK, QPSK, or QAM. Raw bits of data are mapped into symbols using a bit-to-symbol map [4]. This map can be as simple as setting a 0 to one voltage and a 1 to another, or it could involve an incremental phase shift and change in amplitude to represent a longer string of bits. For each symbol period, the phase of the carrier is be adjusted between two values, depending upon the binary data. This is phase shift keying (PSK), and can be done in binary (BPSK) form (single bit per symbol) or quadrature (QPSK) form (2 bits per symbol) [1] [4].

In binary phase shift keying (BPSK), a mixer, driven by the baseband binary data, which provide a +1/-1 modulation, provides either a 0 or 180 degree phase shift depend-

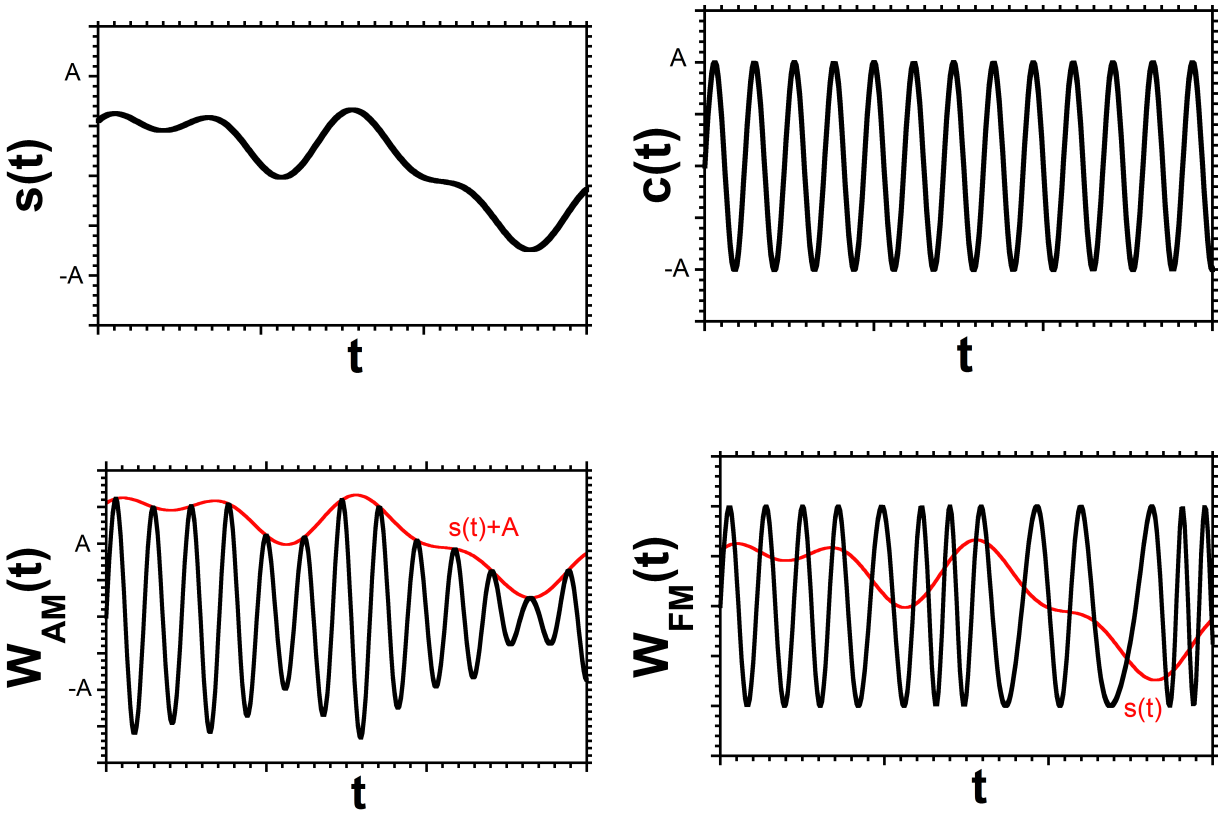

Figure 1.10: Arbitrary sound waveform (top left) mixed with single-frequency sinusoidal carrier waveform (top right) in AM (bottom left) and FM (bottom right) form

ing on the data input [4]. This results in a gain of A or $-A$, which can be mapped to the bits 0 and 1 respectively. A block diagram of a QPSK quadrature vector modulator can be seen in figure 1.11. There are two binary bit streams, $V_{b,I}(t)$ and $V_{b,Q}(t)$. Each of these is mixed with a corresponding carrier signal $V_{LO,I}(t) = V_{LO}\cos(\omega_c t)$ or $V_{LO,Q}(t) = V_{LO}\sin(\omega_c t)$. The outputs are added together to create an RF signal of $V_{RF}(t) = V_{LO}V_{b,I}(t)\cos(\omega_c t) + V_{LO}V_{b,Q}(t)\sin(\omega_c t)$. Since these two bit streams are modulated into carrier signals that are 90 degrees out of phase from one another, they are orthogonal, and can downconverted and separated into the two original binary bitstreams using a similar IQ downconversion vector modulator. Two bits are therefore encoded into each symbol that is transmitted. An example BPSK and QPSK constellation diagram

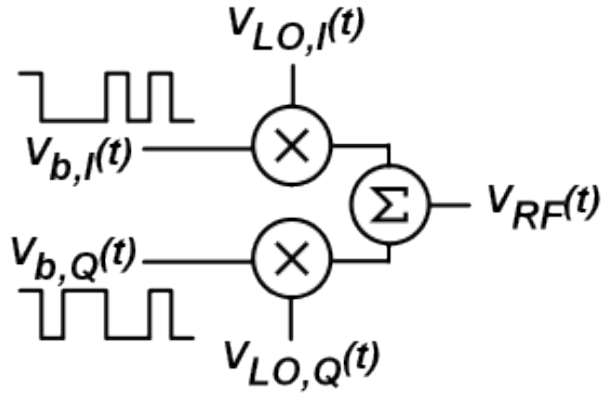

Figure 1.11: QPSK upconversion quadrature vector modulator block diagram

are shown in figure 1.12 [4].

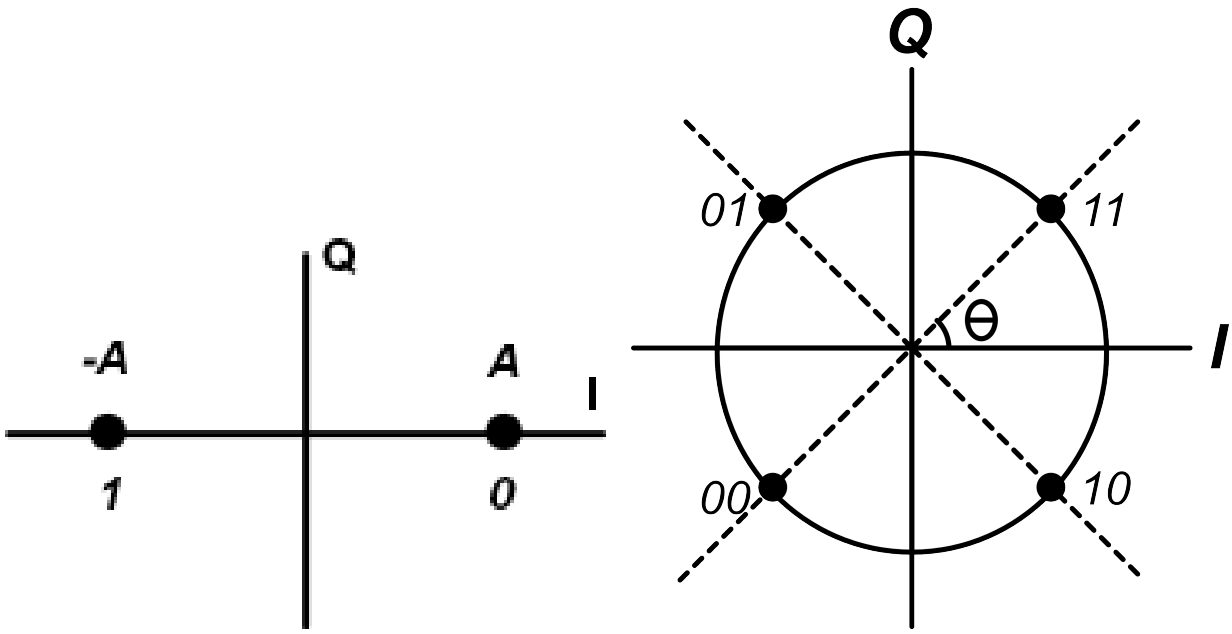

Figure 1.12: BPSK (left) and QPSK (right) constellation diagrams

The number of bits in each symbol can be increased, which creates a more complex constellation pattern. One can do this by applying an N-bit digital to analog converter to both the I and the Q signals, mixing the Q signal with a sine wave of the carrier frequency and the I signal with a cosine wave from the carrier frequency, and then adding these

two outputs. Now the RF signal carries two orthogonal data sets (90 degrees out of phase) with variable amplitude. This is quadrature amplitude modulation or QAM. An example constellation diagram of 16QAM (16 possible symbol values or 4 bits per symbol) can be seen in figure 1.13. The drawback is that a larger number of bits per symbol requires a larger amount of received power per bit of information, i.e. a larger SNR. This is reflected in Shannon's expression, as a more complex constellation provides a greater channel capacity in a fixed channel bandwidth. Noise brings in a degree of uncertainty in both the amplitudes of the I and Q components of the signal. These I/Q constellations represent the received signal in the absence of noise or other receiver impairments. The scale of I/Q constellations are usually normalized such that the radius of the signal vector, i.e. the point of the (I,Q) plane, corresponds to the square root of the signal energy received per modulation symbol, or, equivalently, the received signal power multiplied by the duration of the modulation symbol period. The receiver noise then perturbs each received signal, in both the I and Q directions, by a deflection whose variance is kTF.

The receiver SNR is the ratio of the square of the sifnal radius squared, averaged over the constellation, to kTF. Clearly the closer each symbol is to adjacent symbols in I/Q space, the larger the chance of error [1] [4]. They can be more widely spaced, but this requires a greater received signal power. Remember that the Shannon Capacity only defines the upper limit on the data rate of a channel, and the optimum modulation scheme is chosen to get the actual data rate close to the Shannon Capacity. Hardware limitations will ultimately dictate the most efficient choice for modulation scheme.

Within the scope of this dissertation, it is not necessary to fully comprehend how advanced modulation schemes can be chosen to optimize for power consumption and maximizing channel capacity. It is, however, necessary to understand how a phase shift can be applied to a signal by breaking it into an I and Q components and applying a

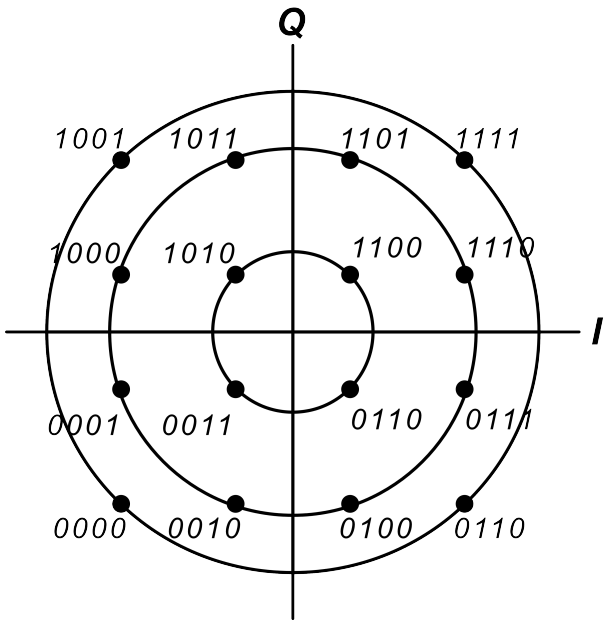

Figure 1.13: Ideal constellation diagram of 16QAM

variable gain to each of them.

Chapter 2

Next-generation Commercial Communications Systems

2.1 Introduction

At-the dawn of wireless telegraphy in the late 1800s, our capabilities were limited by technology - diodes, transistors, and even vacuum tubes had not yet been invented. The development of various technologies has since transformed our capacity to instantaneously transfer information over distances from a modern marvel to a universal necessity in the modern world. Wirelessly transmitting a byte of data (8 binary bits), an implausible idea at the time of Samuel Morse, is now much more than an everyday occurrence. It's projected by Cisco that by 2020, roughly 30.6 exabytes (30.6×10^{18} bytes) will be requested wirelessly through mobile phones and other devices every single month (figure 2.1) [Cisco VNI Mobile 2017]. That is an order of magnitude higher than the 3.7 exabytes requested monthly in 2015. Much of this is attributed to the explosion of mobile high-resolution video streaming. The development of wireless technologies and the consumer demand for them have been feeding off of one another for the past

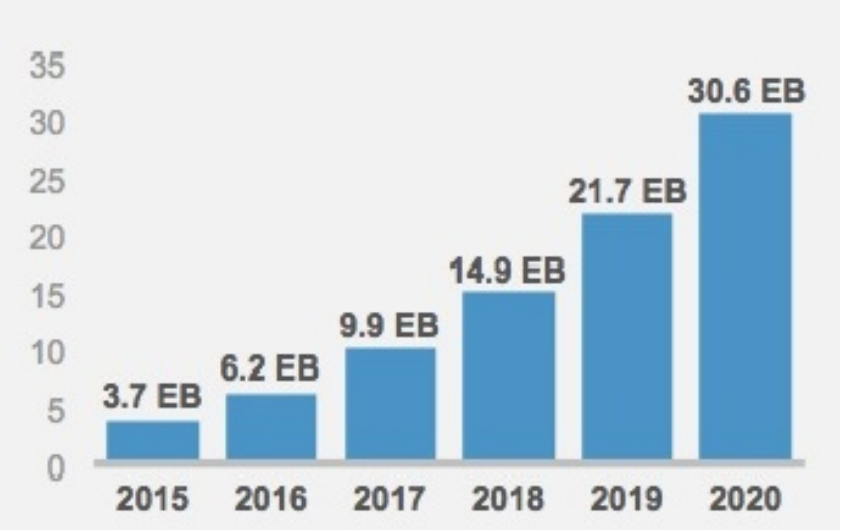

Figure 2.1: Projected monthly wireless data demands worldwide (exabytes per month)

century. As technologies develop to meet demand, it enables new applications, which increases demand further. Even as cellular service providers prepare to implement the 5th generation of mobile infrastructure (5G), we must anticipate further increasing demand and look at more avenues for wireless data expansion by expanding speed in a given channel and by increasing channel density in a given bandwidth.

As technologies develop and bandwidth demands on commercial infrastructure continues to grow, FCC-allocated frequency bands are rearranged to adapt. Increased demands on speed and network bandwidth have prompted the FCC to open up frequency bands at 27.5-29.5 GHz, 37-40.5 GHz, 47.2 - 50.2 GHz, 50.4-52.6 GHz, and 59.3 - 71 GHz to be used for 5G studies and research [FCC 2015]. As of now, they have not yet committed to specific frequency bands for 5G standard. These frequency bands are not only wider than those that were previously used for mobile systems, but they are also at a notably higher frequency than the 2.4 GHz range that was used for the previous generation. In this section, I will look beyond 5G and examine architectures and the design and characterization of circuit blocks which can be used for future generations of commercial

wireless links. The work presented here includes the design and characterization of a low-power scalable multi-channel beamforming matrix which can be used to spatially multiplex multiple channels within a single frequency for high-capacity communications networks.

2.2 Frequency Expansion and Performance Tradeoffs

One direction commercial mobile systems can go to increase network capacity is to increase the network bandwidth. Unfortunately, this is limited in several ways. First, the free space path loss of a propagating beam in the absence of atmospheric attenuation is directly proportional to the square of frequency, so if link frequency doubles, the reliable range is reduced to $\frac{1}{4}$ the distance; given the increased atmospheric attenuation at higher frequencies, the range will become even shorter. Second, it is more difficult to achieve high-performance circuitry at higher frequencies. Power consumption and noise figure will increase and transmitted output power will be reduced. Third, the FCC must allocate frequency bands for many applications with only a finite usable bandwidth, so to open up a broader frequency range for future commercial systems, the wireless industry as a whole must develop high-efficiency, high-performance circuitry to unlock more bandwidth. In this section, I will discuss some of the challenges and tradeoffs involved in designing hardware at higher frequencies.

To generate signals at extremely high frequencies, one method is to generate signals at low frequencies and then to generate the higher frequency signal by harmonic generation using a nonlinear element. This is called frequency multiplication. [15]. The conversion loss for a passive mm-wave multiplier is large, so for most applications, a large amount

of power must be generated at a low frequency before frequency multiplication [15]. An active multiplier can be used which provides gain, but these also have low DC-to-RF efficiency. Producing large AC power at very high frequencies directly is also inefficient [16]. State-of-the-art solid state transistor power amplifiers above 200 GHz can produce an output power on the order of 200 mW, but they do so with power added efficiency below 3% [16].

One reason for the poor performance at high frequencies, is that the maximum stable power gain for a single-transistor amplifier is greatly reduced as frequency is increased [5]. For any transistor technology, there is a maximum frequency of unity power gain, called f_{max}, at which the maximum available power gain and the unilateral power gain are both unity. At frequencies approaching f_{max}, the transistor available power gain is low, and may not be much greater than the resistive losses associated with the transmssion-line impedance-matching networks. Transistor amplifier design at greater than 75% of f_{max} is therefore difficult. The maximum stable gain and Mason's unilateral gain for a 130 nm InP HBT from Teledyne Scientific Company is shown in figure 2.2 [17].

To make things even more difficult, it is in general not possible to simultaneously optimize a design for high power output and for maximum gain or for minimum noise figure and maximum gain [5]. The load impedance for maximum P_{sat} and maximum gain are different, and the source impedance for minimum noise figure and maximum gain are different.

In addition to increased atmospheric attenuation and increased free space path loss at high frequencies, the effects of multi-path fading are also increased at higher frequencies, especially when considering mobile phone applications [4]. This occurs when the transmitted wave takes multiple paths to the receiver, potentially bouncing off of multiple stationary or moving objects. If the line-of-sight (LOS) and non-line-of-sight (NLOS) signals arrive at the receiver during the same symbol period, the signals can interfere

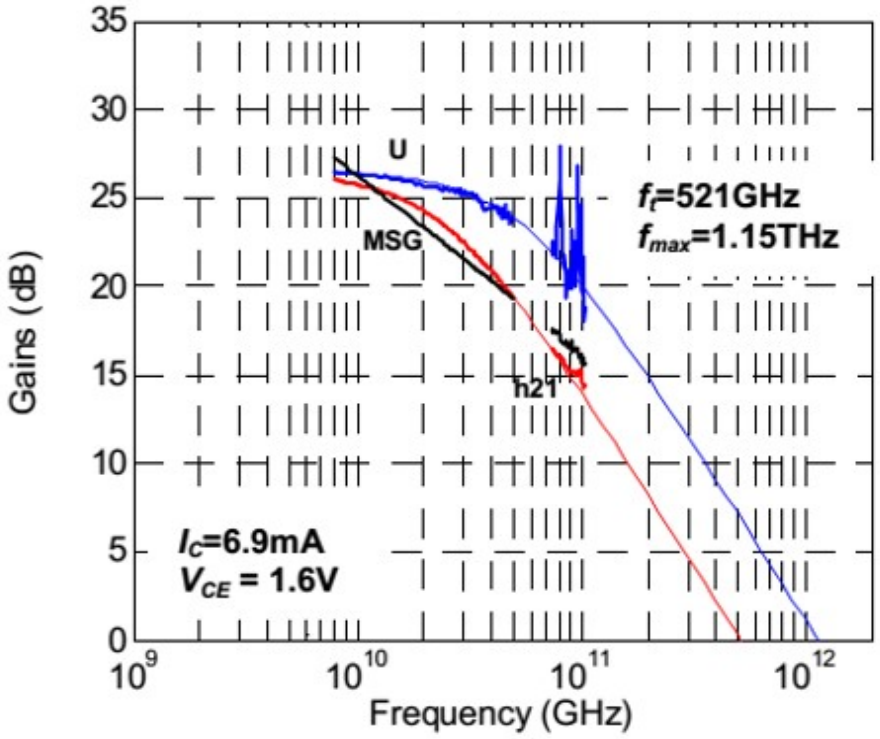

Figure 2.2: Maximum stable gain and Mason's unilateral gain vs. frequency for Teledyne's 130 nm InP HBT process

destructively, causing reduced received signal power. If, on the other hand, the LOS and NLOS signals arrive in different signal periods, then the signals between successive periods will interfere. This is called intersymbol interference. At extremely high frequencies, the beams have a small 1st Fresnel zone, so the LOS signal path is readily blocked. As will be explained in the following sections, the spectral efficiency of these high frequency bands can be greatly expanded using line-of-sight point-to-point spatially-multiplexed low-power mm-Wave communications systems for this type of application [18] [19] [20].

2.3 Silicon on Insulator Technology

Companies in the commercial market are aiming to maximize their profits. Therefore, it is important that they are able to produce a massive number of units for as low a price

as possible. Silicon has been scaled aggressively for computing and VLSI applications for decades, so highly scaled silicon technologies are cheaply available for other markets such as wireless. Thanks to nearly 50 years of aggressive transistor scaling for VLSI applications, there is an extremely mature infrastructure for silicon CMOS technologies. The issue is that silicon technology does not feature as large of an f_{max} as some other material systems, such as InP [13]. Bipolar transistor technologies generally have superior switching speed due to reduced capacitances, so they often are the optimal choice for high-frequency RF applications from a performance standpoint, but are power inefficient for logic and VLSI applications. Bipolar transistor technologies are developed for a smaller volume of customers, and therefore are more expensive for each customer [17].

The key driving force for the development of silicon technologies has been digital VLSI applications. Scaling the dimensions of transistors downwards increases the complexity level of the circuits, allows for lower supply voltages (and lower power consumption), and increases the cost efficiency for large-scale production. There is an added benefit that the reduction in C_{GS} and parasitic access resistances also increases both transistor f_T and f_{max} [21]. Unfortunately, the scaling limits of silicon CMOS technology are getting closer as channel lengths/thicknesses approach the dimensions of individual atoms. Furthermore, the performance benefits of scaling for VLSI applications are starting to be overcome by short-channel effects. Therefore, device engineers have come up with technology innovations to marginally extend performance [21]. One of these innovations is called silicon on insulator (SOI) technology. A cross sectional comparison of a classical planar n-type MOSFET is compared to an n-type SOI MOSFET in figure 2.3. This was developed for VLSI technology for multiple reasons, including reduced short-channel effects, higher resistance to latch-up due to improved inter-device isolation, and reduced power supply voltages. This buried oxide layer also benefits wireless/RF applications through reduced drain-bulk and source-bulk capacitances, which allows faster switching

speeds [22] [23].

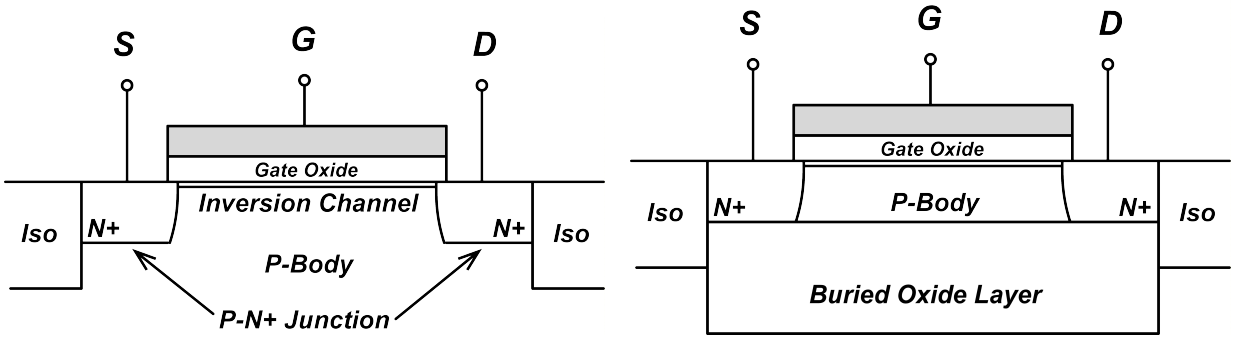

Figure 2.3: Standard FET (left) and SOI FET (right) cross-section

One downside of SOI device technology is the floating body effect. This is the negative consequence of complete isolation between the transistor and the substrate [23]. Since the body of the transistor is isolated from the substrate, holes created by impact ionization under a large electric field in the transistor body tend to move towards the region of lowest potential (the p-type floating body) while the electrons created are quickly swept into the drain. The holes therefore accumulate and build up positive charge. This phenomenon leads to several negative consequences including the kink-effect, potential single-transistor instabilities, negative device conductance and trans-conductance, preventing the transistor from turning off, and premature transistor breakdown [24]. This effect can be avoided by using body-contacted devices for every transistor, but this generally requires larger device widths and potentially increased power consumption depending on the application.

2.4 Phased Arrays and Beamforming

At higher frequencies, free space path loss increases and circuit performance degrades. To overcome these obstacles without drastically increasing power consumption,

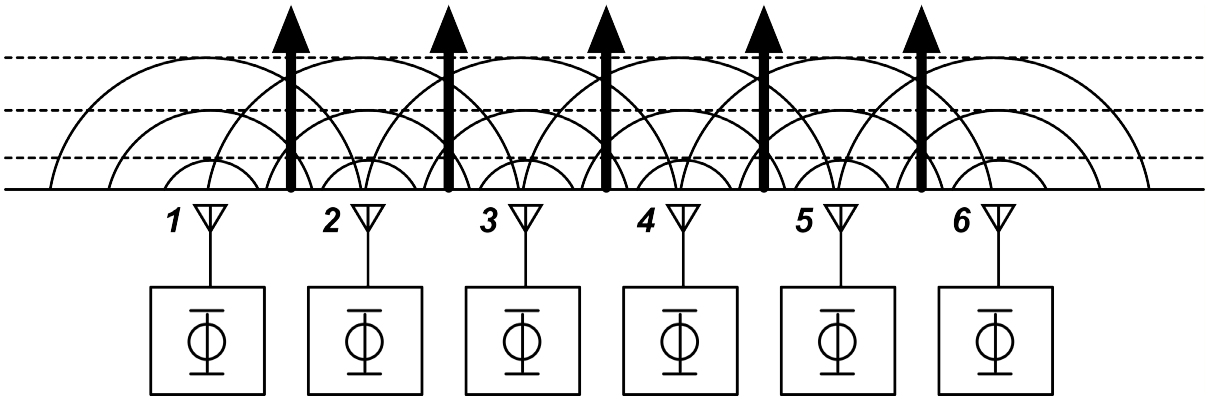

Figure 2.4: 6-element Phased antenna array with no relative phase shift

one method is to use more directional antennas. As opposed to omnidirectional antennas, which are optimum for broadcasting applications, directional antennas steer a beam to a particular location, reducing the amount of power wasted by sending it in the wrong direction. For point-to-point communications, this is the optimum solution. The issue for mobile phones, however, is that one cannot predict exactly where they will be as the antenna is constructed, and mechanically steering a directional antenna is not an efficient strategy either. Therefore, one technique that is now expected to be used for the coming 5G cellular network is to use electrically steerable antenna arrays, called phased arrays. This section will outline the utility of phased arrays and how electrically forming a dynamically steerable beam in this way can increase the effective directionality of an antenna without any unwieldy mechanical hardware or apparatus.

First, see figure 2.4.

First, see figure 2.4. In the figure, there are 6 antenna elements in a linear one-dimensional array. Each antenna element has an electronically controlled RF phase-shift, Φ. First, consider the case where Φ for each antenna element is identical. In this case, every antenna radiates the signal in unison as in figure 2.4. The planes of constant phase (the equiphase fronts) are horizontal along the page, and, effectively, a single beam is formed traveling upwards along the page, or normal to the equiphase front. Much of

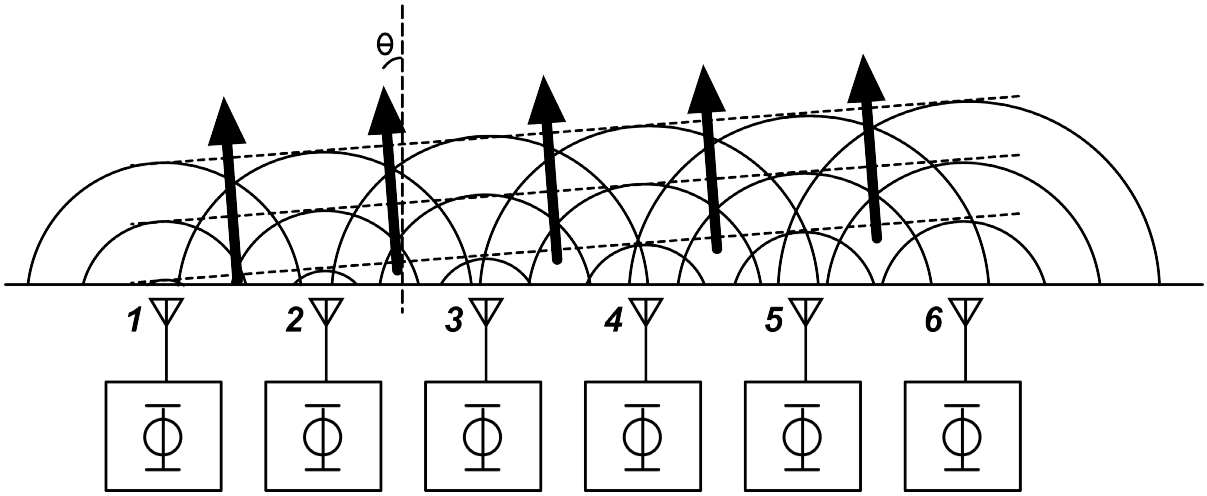

Figure 2.5: 6-element Phased antenna array with relative phase shift

the power that is radiated outwards in unwanted directions is canceled out by varying levels of sinusoidal destructive interference. There is also the added benefit that phased array antennas are bi-directional. In other words, they are equally effective in a receiver as they are in a transmitter, thus doubling their benefit within the Friis transmission equation if used at both ends of the link [7]. By adding a constant increment of phase shift between each adjacent antenna element, the beam can be steered. This technique is called beamforming. If we wanted to transmit or detect a beam at angle Θ as shown in the figure, the difference in phase shift needed between each adjacent element, $\Delta\Phi$, can be seen in equation 2.1 [25].

$$\Delta\Phi = \frac{2\pi}{\lambda} d sin\Theta \tag{2.1}$$

The maximum directivity of the antenna array is determined by the number of antenna elements, the arrangement of the antennas (linear or planar), and also the directivity of each antenna element. Assuming each antenna element is equally spaced and has the same directivity, the directivity of the array D_{array} can be expressed as the product of the directivity of each array element, D, and the array factor, AF, which is the number

of elements in the array [25], (equation 2.2) [25].

$$D_{array} = AF \cdot D \tag{2.2}$$

For arrays steerable over a hemisphere, the element spacing is set at $\frac{\lambda}{2}$. [25]. At the circuit level, a phase shifter can take on a few different forms. The phase shift can be applied in the signal path at the RF frequency or can be applied in the digital baseband or in the LO signal path [26]. Figure 2.6 shows a simplified block diagram of phase shifting in the RF signal path. The phase shifters can be either active or passive [27] [28]. If the wavelength of the signal is small, passive phase shifters consisting of simple transmission lines of different lengths and RF switches to control the signal path can be laid out without consuming too much area on the IC. In the figure, the delay time for each passive phase shifter, T_{Dn}, is approximately equal to $\frac{e^{j\phi n}}{\omega_c}$ [26]. In this architecture, the LNAs at the front of each RF signal path are equipped with variable gain for amplitude control. This architecture has the advantage of reduced dynamic range requirements for the IF and baseband components, since unwanted interference signals incident from directions different from that at which the array is aimed can be canceled after the signals are combined [26].

The phase shift can also be applied in the LO path. The downside is that the dynamic range of the mixers must be larger than in RF signal path phase shifting architectures, since cancellation of off-direction interfering signals does not occur until after the mixers [26]. The RF phase shift can also be applied within the digital baseband circuitry using active phase shifters, like in the architecture shown in figure 2.7. This architecture faces similar but more severe weaknesses as the LO-path phase shifting architecture, in that off-aiming direction interfering signals are not cancelled until the digital signal processor. High dynamic range is thus required for the RF front-end mixers, the IF (if

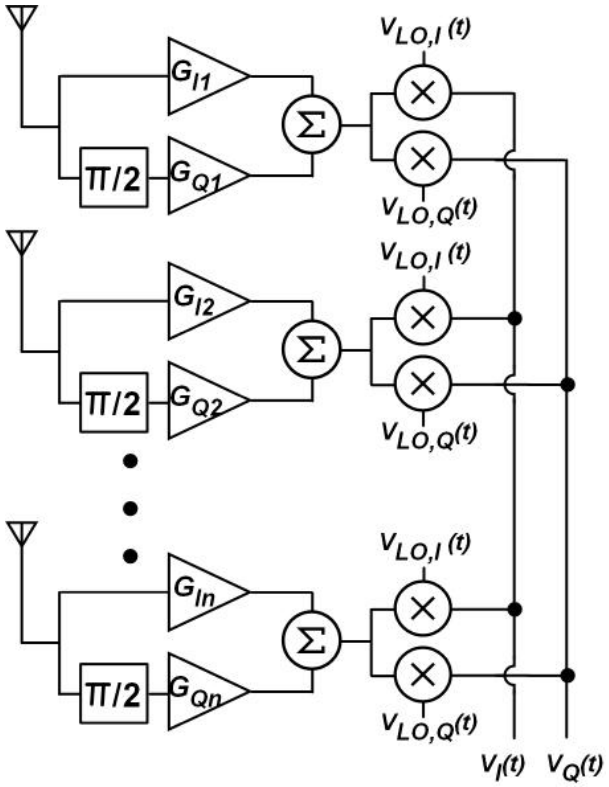

Figure 2.6: Simplified block diagram for a phased array with phase shifting in the RF signal path

present), and the baseband and ADCs [26]. Applying the RF phase shift with baseband circuitry also requires an I/Q downconversion mixer and a pair (I,Q) of ADCs per array element, which consumes more space and potentially more power as well [26]. The key advantage of phase shifting using digital baseband circuitry is that it offers flexibility for applications such as spatial multiplexing and MIMO (multiple-input multiple-output), which will be addressed in the next section.

The concept behind digitally controlled active phase shifters is similar to the concept behind quadrature modulation. In figure 2.6, the QPSK RF signal, $V_i(t)$, is received by each antenna element, where the signal is split into two paths. The first path (I) is sent to a variable gain amplifier (VGA) with gain G_{In}. The second path (Q) is given a phase shift of $\frac{\pi}{2}$ and then to a VGA with a gain of G_{Qn}. These signals are added

together to produce the time-dependent signal $V_{o,n}(t) = G_{In}V_i(t) + e^{-j\pi/2}G_{Qn}V_i(t)$. We can express the complex gain from $V_i(t)$ to $V_o(t)$ to be $G_c = \frac{V_{o,n}(t)}{V_i(t)} = G_{In} - jG_{Qn}$. G_c has a magnitude of $|G_c| = \sqrt{G_{In}^2 + G_{Qn}^2}$, and an RF phase shift of $\phi = -\arctan\frac{G_{Qn}}{G_{In}}$. G_c is therefore expressed as $G_c = \sqrt{G_{In}^2 + G_{Qn}^2}e^{j\phi}$. In this way, the RF phase shift ϕ can be adjusted by changing the values of G_{In} and G_{Qn}. Subsequently, the signal undergoes I/Q downconversion and the outputs are applied to a common load in parallel to extract the appropriate bits from the transmitted I and Q bitstreams. This will result in separate baseband signals $K_I \times V_I(t)$ and $K_Q \times V_Q(t)$ where the K variables are scaling constants dependent on the gain settings used for phase shifting. Since the magnitude of $Vo,n(t)$ for each antenna element depends on G_{In} and G_{Qn}, a VGA would be desirable between the summing block and the I/Q downconversion blocks for QAM modulation formats so that G_{In} and G_{Qn} are not limited to settings which provide a certain $|G_c|$.

Now consider the architecture in figure 2.7, which uses baseband circuitry to apply the RF phase shift at each antenna element. Like in the last example, each antenna element receives an RF signal with a phase shift. In this case, we will express the RF signal at one particular antenna element as $V_{RF}(t) = V_I(t)\cos(\omega_c t + \phi) + V_Q(t)\sin(\omega_c t + \phi)$. The identities $\cos(A+B) = \cos A\cos B - \sin A\sin B$ and $\sin(A+B) = \sin A\cos B + \cos A\sin B$ are useful to give us the expression for the received RF signal, $V_{RX}(t)$. This is $V_{R,I}(t) = V_I(t)\cos(\omega_c t)\cos(\phi) - V_I(t)\sin(\omega_c t)\sin(\phi) + V_Q(t)\sin(\omega_c t)\cos(\phi) + V_Q(t)\cos(\omega_c t)\sin(\phi)$. Unlike last time, I/Q downconversion is applied first, where $V_{LO,I}(t) = \cos(\omega_c t)$ and $V_{LO,Q}(t) = -\sin(\omega_c t)$, and the term with double the carrier frequency is filtered out. Once the terms at double the carrier frequency are filtered out, the expressions for the received I and Q baseband signals are $V_{R,I}(t) = \frac{1}{2}[V_I(t)\cos\phi + V_Q(t)\sin\phi]$ and $V_{R,Q}(t) = \frac{1}{2}[-V_I(t)\sin\phi + V_Q(t)\cos\phi]$.

These baseband signals are applied to the baseband Σ matrix processing circuits. The block diagram of one of these is shown in figure 2.7 on the right. By setting $\frac{G_{21}}{G_{11}} = -\tan\phi$

and $\frac{G_{22}}{G_{12}} = +\tan\phi$, I_{out} and Q_{out} from each Σ block will again be $K_I \times V_I(t)$ and $K_Q \times V_Q(t)$ depending on what the magnitude of the complex gain coefficient is. One note is that the conditions $G_{21} = -G_{12}$ and $G_{11} = G_{22}$ should always be satisfied, as it is always desired to apply the same phase shift to the I component as is applied to the Q component. In the absense of hardware limitations and IC parasitic effects, an RF phase shift can effectively be applied to a signal using baseband circuitry and the result is equivalent to that of active RF signal path phase shifting architectures.

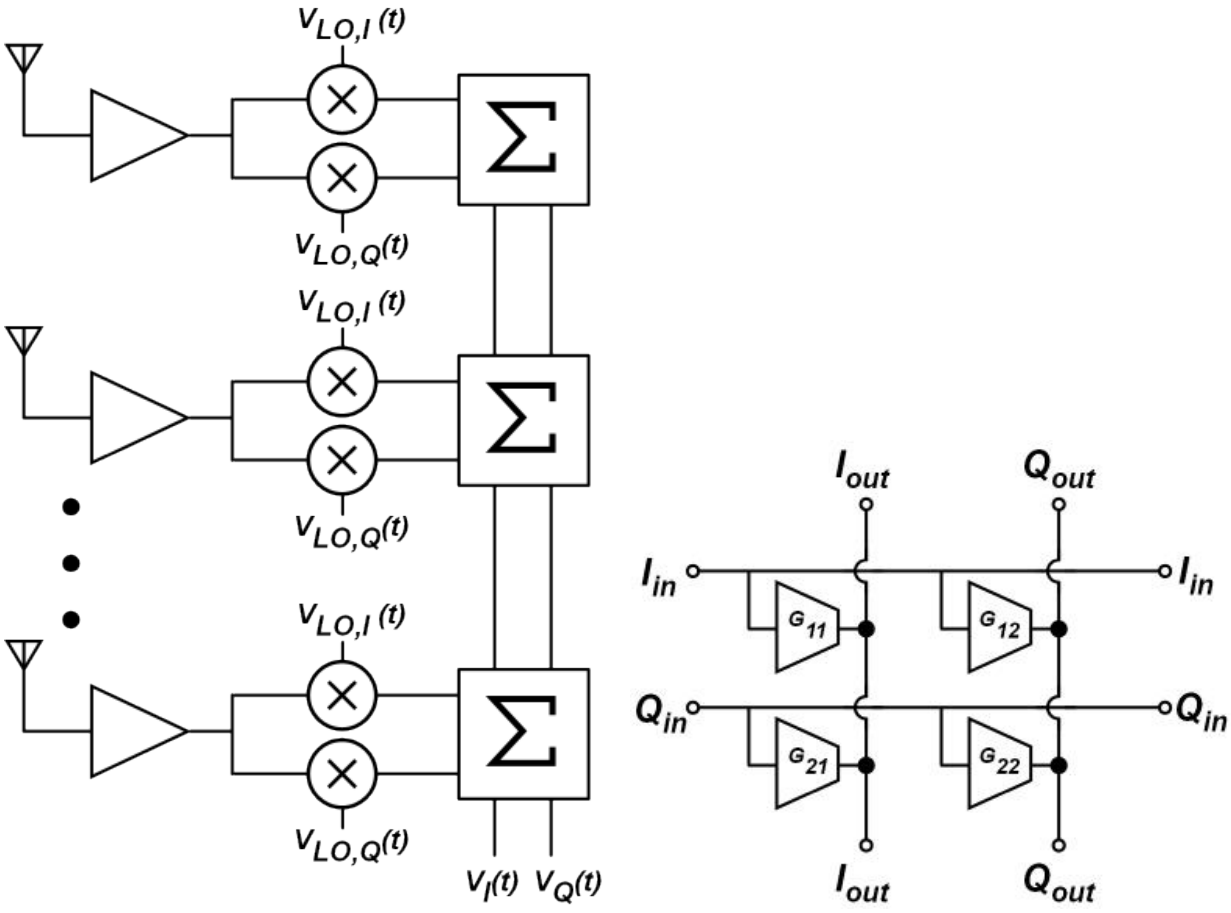

Figure 2.7: Simplified block diagram for phased array architecture with RF phase shifting in the digital baseband circuitry (left), and block diagram of baseband Σ matrix processing circuit (right)

Arrays can be arranged in a linear row and steer the beam in a single lateral direction

as shown in the figure, but they are also often designed in a planar topology, which enables them to steer in 2 directions, azimuth and elevation, and allows for higher directivity with the expense of design complexity [25].

2.5 Spatial Multiplexing

Phased antenna arrays can achieve higher antenna gains using electronic steering of a highly directional beam. This opens the door to much higher RF frequencies with a larger SNR over somewhat short distances with a line-of-sight. They can also increase the number of local wireless channels within a particular bandwidth using spatial multiplexing when the the receiver or transmitter design is expanded to separate independent beams from different angles.

As shown in the previous chapter, the Shannon Capacity for a particular channel is dependent on the bandwidth and SNR. The overall network capacity with n channels within a particular finite bandwidth is therefore shown in equation, where the sum of BW_i is the overall bandwidth of the network 2.3 [11] [25].

$$C_{network} = \sum_{i=1}^{n} BW_i \log_2 (1 + SNR_i) \tag{2.3}$$

Most of the time, channels are separated by frequency multiplexing. Woth each channel centered at a slightly different frequency and offset in frequency so that the interference and power leakage between adjacent channels is limited [1]. The SNR of each individual channel will depend upon the location of that particular handset relative to the tower, hence it is difficult to calculate the capacity of a general network. If we assume that every handset is optimally placed such that it has a maximum possible SNR,

defined as SNR_{max}, then the capacity becomes

$$C_{limit,network} = BW_{total}\log_2(1 + SNR_{max}) \tag{2.4}$$

This is the fundamental limit of data rate for a frequency-multiplexed network with a total bandwidth of BW_{total} and a maximum SNR of SNR_{max} on each channel, but the actual capacity will be below this because frequency guard bands are needed to prevent adjacent channel interference, and because not all handsets will receive the maximum signal power [1]. If instead we allow re-use of frequency spectrum using spatial multiplexing, the overall network capacity limit becomes multiplied by the number of channels co-existing in each frequency. If a base-station employing a multi-beam phased array system for both transmit and receive modes allows for m separate beams in the same frequency, the overall local network capacity is effectively multiplied by the number of beams at each frequency band used.

By using a highly-scaled low-power multi-beam phased array system, spectral efficiency could be greatly increased [18] [19] [20] [29]. MIMO base stations of this type could be implemented outdoors on busy streets, plazas, or city centers in populated urban areas or at event locations like stadiums or exposition centers. The number of independent beams is the number of elements in the array. The elements must have a spacing of $\frac{\lambda}{2}$ for steering over a hemisphere, hence the array area needs to be $A = N \times \left(\frac{\lambda}{2}\right)^2$ where N is the number of beams. Higher frequencies permit smaller area arrays for a given number of users, but lower frequencies suffer less from beam blockage and will have a longer range due to lower atmospheric attenuation and lower free space propagation loss [7].

Phased arrays can greatly increase network capacity by using spatial multiplexing to greatly boost spectral efficiency. Multiple channels can exist on the same frequency

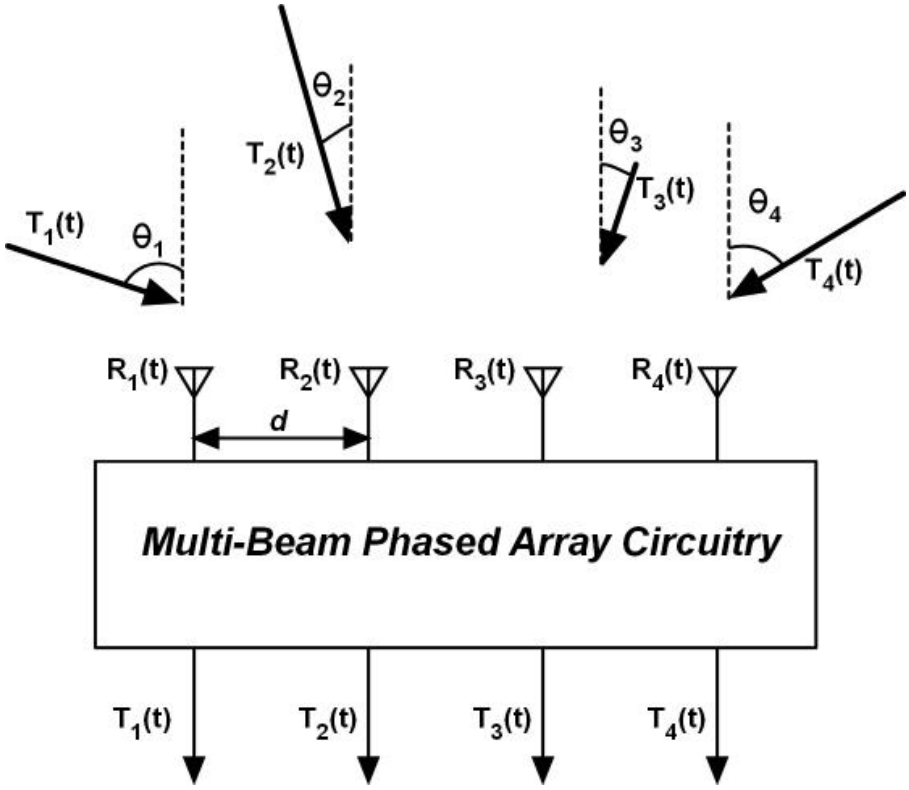

Figure 2.8: Multi-beam phased array receiver receiving 4 signals at the same frequency simultaneously

band. Figure 2.8 shows spatial multiplexing of 4 signals all received at the same frequency channel but coming from different angles with different powers. The transmitted signals for each received channel are labeled $T_1(t)$ through $T_4(t)$. We will call the signals received by each antenna element $R_1(t)$ through $R_4(t)$. Each $R(t)$ is some linear combination of each $T(t)$ with phase shifts between adjacent antenna elements. Since we are dealing in relative phase shifts, we can define the initial phase shift at the first antenna element to be 0 degrees. For each spatially multiplexed channel, the phase shift between adjacent antenna elements is $\Delta\Phi_n = \frac{2\pi}{\lambda} dsin\Theta_n$. Therefore, the vector of antenna inputs can be

represented as follows

$$\begin{bmatrix} R_1(t) \\ R_2(t) \\ R_3(t) \\ R_4(t) \end{bmatrix} = \begin{bmatrix} X_1 & X_2 & X_3 & X_4 \\ X_1 e^{-\Phi_1} & X_2 e^{-\Phi_2} & X_3 e^{+\Phi_3} & X_4 e^{+\Phi_4} \\ X_1 e^{-2j\Phi_1} & X_2 e^{-2j\Phi_2} & X_3 e^{+2j\Phi_3} & X_4 e^{+2j\Phi_4} \\ X_1 e^{-3j\Phi_1} & X_2 e^{-3j\Phi_2} & X_3 e^{+3j\Phi_6} & X_4 e^{+3j\Phi_4} \end{bmatrix} \begin{bmatrix} T_1(t) \\ T_2(t) \\ T_3(t) \\ T_4(t) \end{bmatrix} \tag{2.5}$$

which can be expressed as

$$\vec{R}(\ t)\ = \mathbf{M}\vec{T}(t) \tag{2.6}$$

By multiplying each side of the equality by the inverse matrix, we get

$$\mathbf{M}^{-1}\vec{R}(t) = \vec{T}(t) \tag{2.7}$$

Where each element in the inverse matrix contains a linear gain coefficient and a phase shift. As long as there is a high level of control over the gain and phase shift from each input antenna to each output channel, and as long as there is sufficient dynamic range, it is possible to independently reconstruct each separately transmitted signal into its own output channel for digital processing. By scaling this up to much larger dimensions - for example, 1000 antenna elements in a planar arrangement - there could be hundreds of channels in a single frequency channel allocation, which could drastically expand the capacity limits of a local wireless network in a small densely populated area.

In reality, this is not a simple task. Any multi-signal monolithic circuit is going to have some finite level of crosstalk between channels, so it is a challenge to separate multiple independent channels on a single chip. This is particularly true when some channels are being received with power levels orders of magnitude higher than other channels, which is common for mobile networks covering a large area. Independent recovery of signals

of very different magnitudes will be limited by the finite resolution and precision of the settings of the gain elements of the matrix, by intermodulation of the signals within the matrix, and by noise. The number of recoverable channels, and the allowable difference in signal power between them will thus depend upon the dynamic range and precision of the beam recovery matrix. Furthermore, as the number of elements and channels scale upwards, the number of on-chip variable gain amplifiers and phase shifters scales up as well, leading to a rise in power consumption. Therefore the proposed system in the following section will be designed with the explicit intention to keep power consumption as low as possible for a given symbol rate.

Chapter 3

Mm-Wave Power Amplifiers and High-linearity Amplifiers

3.1 Introduction

One of the two factors which determine channel capacity is bandwidth [11]. Typically, this is a factor which is out of the control of the circuit designer, as it is dependent on the frequency space allocations designated by the FCC. However, periodically, the FCC reorganizes the allocated sections of bandwidth, expanding the bandwidth for some applications and reducing it for others, and allocating higher frequencies as technology permits. The military, unrestricted by the government allocations, sponsors the design of wireless equipment for radar and imaging and electronic warfare at much higher frequencies than what is typically done for commercial applications. Ultimately, the performance of equipment designed at these frequency bands determines what the maximum achievable performance is at high frequencies and also determines roughly the maximum feasible bandwidth for wireless applications given a particular process technology. One major advantage of military-grade technology is that it is often not necessary to make

it feasible for mass production or affordable for the average consumer. This opens the door to state-of-the-art InP-based HBT process technologies which feature maximum unity power gain frequencies larger than 1 THz. The military will also often be willing to incorporate circuitry with lower power efficiencies than would be used in the commercial sector to achieve functional levels of performance at frequencies unreachable by competitors.

The challenge is to improve performance and power efficiency at higher frequencies. At the device level, the maximum possible achievable gain is low when the target design frequency is larger than 10 percent of f_{max}. For an amplifier designer, this limits the designer's ability to optimize for power output, power efficiency, linearity, and noise. As explained in chapter 1, optimizing matching impedances for any of these features sacrifices gain. If a designer sacrifices gain for output power at a high frequency, they may achieve a high maximum output power, but it's not helpful if it cannot simultaneously provide significant gain. Similarly, a low noise amplifier does not limit the system noise figure if it does not provide a significant gain as well. In this way, the maximum device speed and bandwidth are the limiting factors for performance at high frequencies.

3.2 Design of Mm-wave Power Amplifiers

Achieving high power output, high linearity, and high power-added efficiency in a transistor amplifier are of critical importance for the design of power amplifiers for high-performance wireless links [5] [1]. The free-space path loss of a high-frequency signal is large, so it is important that the transmitted signal is sent with as much power as possible so that it can achieve maximum range with low distortion. If using a passive mixer, the LO driver amplifier will need a high saturated output power to improve the mixer IP3[1]. In this section, I will specifically outline the design considerations for

mm-wave power amplifiers. This tutorial will focus primarily on bias point amplifiers, particularly class-A and class-B, as higher order amplifiers require careful consideration of impedance transformations of harmonic frequencies, which becomes unwieldy as the design frequency increases. Furthermore, the higher gain offered by class-A and class B amplifiers is often more valuable at mm-wave frequencies than the increased collector efficiency offered by higher-order amplifier types.

The first, and arguably most important, considerations are the choice of size and arrangement of the transistor power cell. Making the correct choices early on in the design process is critical, as it defines the ultimate limitations on the amplifier gain, output power, and efficiency. It is impossible to maximize all of these factors simultaneously [1] [5], so decisions must be made based on the system requirements. It is typical that the system will require a certain amount of power at either the saturated output power condition or the 1-dB compressed gain condition starting from a lower amount of input power [1]. In this case, the goal is to achieve this output power with as much efficiency as possible, and enough gain to raise the input power level to the desired output power level [1]. It is also important to consider whether a common emitter/source topology is better or a common base/gate topology. At high frequencies, common-emitter stages typically have lower gain than a common-base stages. The downside of common-base stages is that they tend to have more stability problems [**?**], which can be difficult to manage at frequencies below 10 percent of f_{max}.

Assuming the designer has no control over the type of technology they are designing in, the first consideration is the safe linear operating regions of the transistor bias conditions. Figure 3.1 shows an example of a safe linear operating region of a common-emitter bipolar transistor transposed on the I-V plane. Notice there is a maximum amount of current, a breakdown voltage, and a minimum voltage in the linear region. Typically, the maximum allowable current is on the order of 4 mA per linear micron of emitter length for a bipolar

transistor and on the order of 0.5 mA per linear micron of gate width for MOSFETs. The safe linear region for voltage goes from the CE or CB breakdown voltage down to the boundary between the linear active region and the non-linear saturation region. For MOSFETs, this corresponds to the boundary between the linear saturation region and the non-linear triode region. For a class-A amplifier, these boundaries determine how much power can be pushed out of each micron of emitter length or gate width, thus determining the minimum total emitter length or gate width that is required to achieve the desired output power with equation 3.1. Dividing the total amount of required power by the maximum amount of power from each micron of emitter length or gate width tells the designer what the absolute minimum amount of total transistor finger length is required.

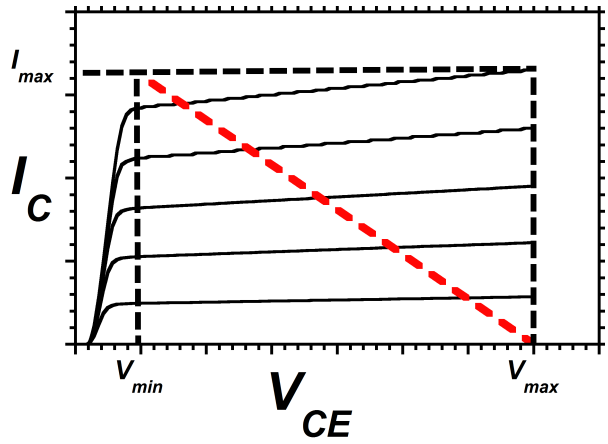

Figure 3.1: Common-Emitter curves for different base biasing conditions. The red dotted line represents ideal load line swing

$$P_{out} = \frac{V_{max} - V_{min}}{8I_{max}} \tag{3.1}$$

In ADS or Cadence, an ideal load line simulation can be performed by running a harmonic balance simulation on the ideal device with a bias current at approximately half of the maximum allowable current density and a bias voltage approximately halfway between the minimum and maximum boundaries of the linear region of operation. The

goal is to determine the appropriate output impedance for a given transistor emitter length or gate width by applying a sinusoidal input signal at the design frequency, and plotting the transient curve of the collector/drain current and collector-emitter/drain-source voltage for full wavelength cycles at various input power levels. Since this is an ideal simulation and there are no additional parasitics, the real part of the output termination should just be set to $\frac{V_{max}-V_{min}}{I_{max}}$ and the imaginary component should be very close to the shunt inductance needed to cancel out the intrinsic transistor output capacitance. If this is done correctly, the transient curve should appear to be a line between the top left corner and the bottom right corner of the linear safe operating region of the transistor. If the slope is incorrect, the real component of the impedance should be adjusted and if the transient curve appears as an ellipse instead of a line, the imaginary component should be adjusted.

At this point, the design begins to get much more complicated at higher frequencies. It is expected that there will be losses and unwanted extrinsic parasitics at the device level which will limit the output power of the realized amplifier. If the designer chooses to design using the absolute minimum emitter length/gate width, then they will certainly come up short of the system requirement for power level. In general, the extrinsic parasitics and combining losses are going to be larger if the amplifier uses a larger total power cell size. Therefore, most mm-wave power amplifier designs require some level of iteration and trial and error to determine the optimum approach. If the designer gives themselves too small a margin of error, it may turn out that it is impossible to actually achieve the system power level requirements. If the designer gives themselves too large of a margin of error, it will reduce the output stage gain and consume more power than necessary, cutting into efficiency and requiring a larger input drive power. This decision is further complicated in that there are multiple ways to divide up this transistor area, and each way will provide different advantages and drawbacks. The full effects of each

method cannot be easily computed by hand. Often the only way forward is to design a set of transistor extrinsic footprints of varying finger lengths, numbers of fingers, and overall sizes, perform electromagnetic simulations, and perform load line simulations for each one.

In choosing transistor sizes within a power amplifier, the designer must consider the length of each transistor finger, the number of fingers in each transistor cell, and the number of cells combined in parallel. Depending on the transistor technology and frequency of operation, the optimum approach may vary significantly. Increasing the length of transistor fingers reduces the total number of fingers needed, but in bipolar transistors, it increases the emitter access resistance and inductance, which hurts the gain and frequency performance. For a MOSFET, this is a little more complicated, as increasing the gate width reduces gate-drain and gate-source overlap capacitance, but increases access resistances and inductances, therefore there is an optimum finger length for a particular frequency and number of fingers in each device.

In determining the number of fingers within a transistor cell, the designer must consider several factors. Of course, the number of fingers in each transistor cell will have a direct impact on extrinsic footprint parasitics. A larger number of fingers is capable of producing a larger amount of power output at a particular level of gain compression, however, the additional parasitic inductance and capacitance from the larger cell reduces the gain and makes the trade off with output power more costly. For designs using MOSFET technologies, each device typically uses a variable number of parallel fingers. In addition to the increase in access parasitics caused by larger devices, the increased width of the device can cause some level of destructive interference or distortion at high frequencies. If the width of an $n-$finger device is large compared to the wavelength, the phase shift of the wave applied to the edge fingers will be notably different from the phase shift applied to the center fingers. If the power from each finger is recombined out of phase, it will

result in destructive interference or harmonic distortion.

In the case of bipolar transistors, the emitter fingers must be well-balanced to prevent current hogging. For example, if there is an odd number of fingers within a multi-finger cell and the finger in the middle sees less access resistance than the other fingers, it will take more current than the other fingers. This causes it to heat up, resulting in an increase in collector current relative to the other transistor fingers. This in turn causes that finger to heat up more, producing a catastrophic self-reinforcing cycle that will destroy the device. To avoid this, the easiest approach is to stick to powers of two with careful symmetrical placement of vias. Even with perfect symmetry, there is a risk of catastrophic current hogging if the fingers are too close together and heat produced by the fingers on the edges is dissipated more easily than for the fingers in the middle of the cell. The only way to be absolutely certain of a safe transistor spacing for bipolar transistor technology is to run a full thermal computer simulation. Without access to this type of simulation or time to carry it out, the designer must resign to a conservatively large spacing between fingers or risk catastrophic thermal failure.

The key goals to aim for when designing a transistor footprint for a power amplifier are to limit the input/output capacitance as much as possible and to limit any emitter/source lead inductance in common emitter/source amplifiers. Similarly, it is important to limit base/gate lead inductance in common base/gate amplifiers. Input/output overlap capacitance is a feedback mechanism which increases the S_{12} parameter. The equations for Maximum available/stable gain in equations 3.2 and 3.3 show that the maximum possible gain is reduced when S_{12} is increased, where K is the Rollet stability factor.

$$MAG = \frac{|S_{21}|}{|S_{12}|} \tag{3.2}$$

$$MSG = \frac{|S_{21}|}{|S_{12}|} \times (K - \sqrt{K^2 - 1}) \; for K > 1 \tag{3.3}$$

Similarly, unwanted emitter/source lead inductance on common emitter or common source amplifiers reduces maximum available gain. At high frequencies, where losses are high, parasitics are dominant, and available gain is low, it is critical to not simply waste potential gain. If base or gate lead inductance is added to a common base or common gate transistor cell, it can cause instability and oscillations.

Once the power cells are designed and have been adequately optimized for the system requirements, it is time to consider the best method for output power combining. There is some room for the designer to get creative at this point as long as they never lose sight of their overall objectives and design principles. At this point, it is a good idea for the designer to run a load line simulation to determine exactly how much impedance should be seen by the output of each transistor cell. The output impedance of the load line should be adjusted to compensate for the change in size of the transistor cell and the added extrinsic parasitic elements. The designer should also check and make sure that the small signal gain with a gain-matched input and output impedance optimized for maximum output power is large enough to be useful. Keeping in mind that adding non-ideal matching networks and power combiners will reduce the stage's realized power gain, the gain likely needs to be at least 6-8 dB to provide any substantial power gain from the stage. To combine 4 or 8 power transistor cells, the gain may need to be even higher than this. At extreme frequencies near 20 percent of f_{max} or higher, it may be that the output impedance needs to be balanced between the optimum gain-matched condition and the impedance for maximum power output. I have found that a good technique for accomplishing this is to plot the optimum impedance for maximum power output and the impedance for maximum operating gain on a Smith chart in a small

signal simulation, draw a straight line between the two points, and test the output power at various impedances along this line until the gain is satisfactory. This is the optimum output impedance that each transistor cell should see at the output.

From this point, the impedance which should be presented to each transistor cell is known and the impedance at the output of the combiner is either the complex conjugate of the input impedance of the following circuit block or the system impedance. The goal is to match to these impedances with as little insertion loss as possible. For a power amplifier, particularly at high frequencies, the performance is degraded in nearly every way by losses at the output of the amplifier. The gain and output power are directly reduced by output losses and the power consumption stays the same, which greatly reduces the power-added-efficiency as well. The power combiners should be symmetrical with binary symmetry if possible. In most cases, extremely high frequency power amplifiers will not be able to work well with classical binary corporate power combiners with more than 8 parallel cells, because the additional insertion loss going from 4:1 to 8:1 combining will likely be on the same order as the boost in output power (3 dB). The power consumption will be twice as high as well.

For pre-stage amplifiers, it is critical to make sure the last stage driving the output stage has enough output power to push the output stage to the required level of output power before it begins saturating. If this is the case, the same technique used to find the impedance balanced between gain matched condition and maximum power output condition can be used to determine the impedance for the output of the final driver stage. The input of each stage can be matched for optimum power gain, and the first few driver stages can be gain-matched on the output as well.

3.3 Mm-wave High-Linearity Amplifier Designs

To a certain extent, a high-frequency high-linearity amplifier is just a slightly modified class-A amplifier. The procedure to determine the optimum input and output impedance is exactly the same in each case, as both circuits aim to maximize IP3 for their particular transistor topologies and bias conditions. In a power amplifier, the designer is optimizing the IP3 of an amplifier to achieve a large amount of output power at a certain level of gain compression. In a high-linearity amplifier, the designer is aiming to achieve a certain level of IP3 with as little power consumption as possible. This type of amplifier is often used after the first mixer in a dual-conversion architecture to improve the linearity of the system, and in military applications, make the receiver more resistant to in-band jamming signals. In most cases, the first IF is a low frequency, which allows for larger levels of linearity with lower power consumption. However, recent developments in high-speed device technology have begun to make it feasible to design high linearity amplifiers at frequencies as high as 100 GHz. This opens the door to broadly tunable architectures that are not possible with a low-frequency IF. In this section, I will discuss the design of mm-wave high-linearity amplifiers and the potential applications for them.

There are many widely-explored techniques for boosting the linearity of an amplifier at low frequencies. This often involves the cancellation of higher order harmonic terms using feedback elements, similar to the methods used for higher order power amplifiers. At high frequencies, the separation between harmonics becomes larger and the matching networks used at the fundamental frequency requires a high level of design precision. This means a tuning or feedback mechanism for higher order harmonics becomes difficult to design without inherently interfering with the design of precise matching networks for the fundamental. Therefore, in designing a high-linearity amplifier at extremely high frequencies, it is preferable to stick to class-A common-emitter/common-source impedance tuning

techniques to optimize linearity and to use inductive transmission line emitter/source degeneration to boost IP3 in exchange for power gain. To do this effectively at higher frequencies, technologies with a very high f_{max} are needed. High linearity amplifiers can be characterized in terms of the ratio between OIP3 and DC power consumption. Since input-referred IP3 can be improved simply by dropping the gain of the amplifier, the output-referred IP3 is a more relevant figure of merit with regards to the quality of the design. Some people like to incorporate frequency into the calculation for figure of merit, however, it is not clearly quantifiable how the difficulty to achieve high IP3 scales with frequency, so I prefer to avoid this type of characterization.

For a high-linearity amplifier used in the IF chain of a wireless system, the designer is usually attempting to achieve either a maximized system dynamic range or some threshold of system linearity. Therefore, in most cases, it is not usually clear what the basic minimum requirements are for a high-linearity amplifier unless the performance of the first conversion mixer is already sufficiently well-known. The best way to design an amplifier to maximize system dynamic range, is to keep a gain budget log with the gain, noise figure, and linearity of each circuit block in the system, enter in projected values throughout the design process, and adjust the design goal to meet the values which produce the best results. This is important because there will be iteration between the amount of inductive emitter degeneration and the gain budget to determine the optimum trade off between gain and linearity.

Just like in any other high frequency amplifier design, the design flow begins with the extrinsic transistor footprint and the transistor cell. In general, a larger transistor cell will result in a higher maximum achievable linearity. This makes sense because a larger amount of current can be pushed out of a larger transistor in class-A operation without pushing towards the edges of the safe linear operating region. The drawback of making the transistor cell arbitrarily large is that it increases parasitics at the output. Losses

in the output matching network cut directly into the linearity with no compensation. The transistor footprint, unlike the power amplifier cells from the previous section, will require an open port at all three device ports. In the power amplifier design, it is critical to preserve as much gain as possible. In the high-linearity design, we would like to trade some gain in exchange for higher linearity by adding some inductive transmission line degeneration on the emitter or source, generally preserving at least 5 dB of power gain or more so that the noise figure of subsequent circuit blocks has less of an impact on the overall dynamic range of the system. The tricky part is that the optimum impedance presented to both the input and output will be different depending on the amount of degeneration.

This can be a tough problem to deal with, because now the optimum size of the transistor cell and the length of source or emitter degeneration determine the optimum impedances to present to the cell input and output, and the performance of the amplifier with the optimum impedances need to be known to determine the best combination of transistor cell size and emitter degeneration. The strategy that I use is to simply design 2-3 different sized 3-port transistor cells and perform an electromagnetic simulation. I place each simulation result within a schematic harmonic balance simulation of the amplifier and apply a bias current on the order of 0.2 - 0.3 mA per micron of emitter length or 0.1 mA per micron of gate width. I place an ideal transmission line to use as the emitter or source degeneration, and then perform a load line simulation to determine the optimum output impedance for maximum linearity. Then I quickly place ideal input and output matching networks with inductors and capacitors to reach the optimum gain matching condition on the input and optimum linearity condition on the output. I perform a two-tone harmonic balance simulation to record the IIP3 and OIP3, and I perform an S-parameter simulation to record the gain and the noise figure. I then plug these values into the gain budget and repeat the process for a few different values of

emitter/source degeneration until I think I have found the length which provides me with the maximum system dynamic range. I do this for each size of transistor cell, ultimately eliminating my options down to one transistor cell and some narrow range of emitter/source inductance. This may seem like a long and inefficient strategy, but determining the optimum transistor size and degeneration inductance is the most critical piece of the design flow, as it determines the limits of the amplifier's performance and the other steps are largely procedural.

One more trick which can be used to boost linearity at high frequencies is to apply a second-harmonic short circuit at the input and output of the transistor cell. For a high-linearity design, the goal is to achieve a large gain at a fundamental design frequency, and suppress the in-band 3rd harmonic distortion. There is also a second harmonic tone which resides at double the fundamental frequency. This signal does not directly contribute to the 3rd order intermodulation product, however, it can mix with the fundamental signal, which is another mechanism for 3rd order distortion. To do this, a $\frac{\lambda_d}{4}$ transmission line with a ground termination through a large capacitor can be used. At the design frequency, this appears to have infinite impedance, and therefore has no negative impact on the matching network or the performance of the amplifier. At the second harmonic which is at double the design frequency, this is $\frac{\lambda_d}{2}$ which appears as a short circuit. Therefore, the second harmonic tone can be eliminated from the output and the input with no drawbacks aside from the additional space that is consumed.

3.4 Tapered-line Distributed Amplifiers

Aside from corporate power combiners, there are other creative ways to combine multiple power cells in phase. One such way is by using a tapered-line distributed amplifier. A distributed amplifier uses transmission line theory to combine the power output

from multiple transistor amplifier cells in phase in two directions to increase the gain-bandwidth product of the amplifier [30]. A tapered-line variation of this works in much the same way, except the reverse-direction output line termination is eliminated. In this section, the concepts of tapered-line distributed amplifier design and the drawbacks and benefits will be outlined.

A circuit diagram of a high-frequency tapered-line distributed amplifier can be seen in figure 3.2 where 4 common-emitter transistor cells are being combined. In this case, if impedances do not naturally match to the correct impedance for maximum power transfer, a matching network represented by an "M" in a box is inserted. The input is fed to a transmission line with characteristic impedance $\frac{Z_o}{4}$. Since these amplifier cells are combined in parallel, the transmission line is carrying a single voltage waveform which will be applied to the input of each cell, which we will simply define as V_{in}. However, the current carried by this section is the combined sum of the current waveforms of each cell. If we define the current waveform being presented to each transistor cell as I_{in}, then this transmission line section is carrying $4I_{in}$. This transmission line section sees a Z_o input impedance in parallel with a $\frac{Z_o}{3}$ transmission line. Therefore one quarter of the current is applied to the input of the first amplifier cell, and the other three quarters continue down the tapered input line. In a similar fashion, each of the following junctions apply a voltage waveform V_{in} and a current waveform I_{in} to each amplifier cell input.

At the output, the same method of transmission line tapering is used to make sure that the signals are being added together in parallel. The output from each amplifier can be seen as a source of current I, with a voltage V. Since, like at the input, these are being combined in parallel, the voltage is V at each node and this does not change as the amplifiers are combined. The currents add. Unlike the input tapered transmission line, however, which was gradually giving off current to each amplifier to avoid seeing a reflection caused by an impedance mismatch, the output line does see reflections at

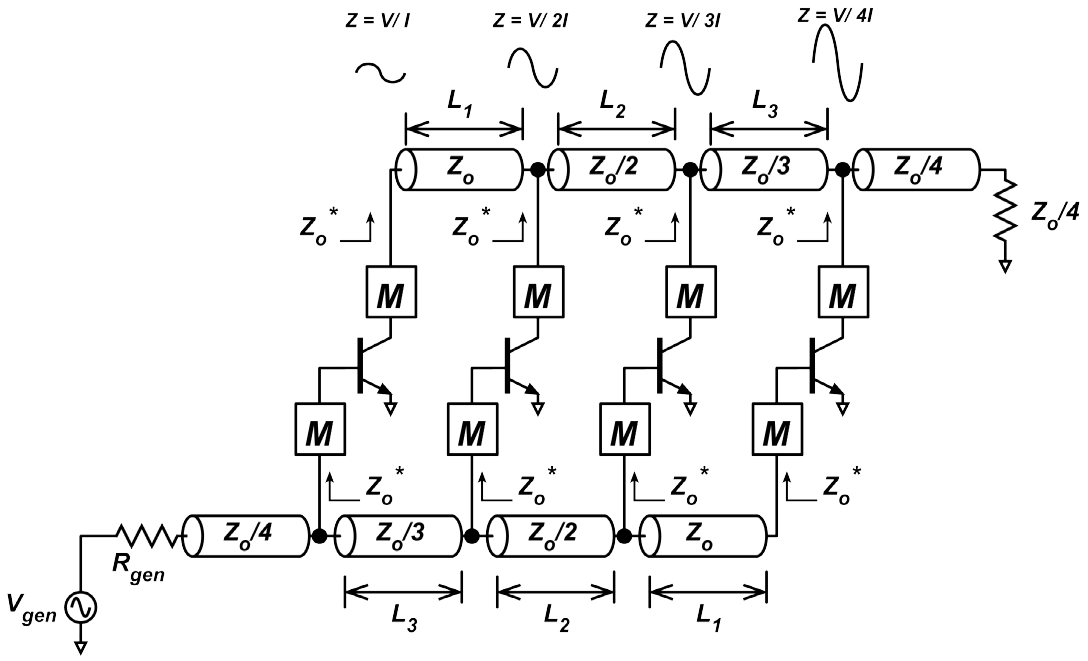

Figure 3.2: 4-cell tapered line distributed amplifier block diagram

these junctions. Fortunately, if the current waveforms from each transistor are tracked and the reflected and transmitted components at each junction are meticulously added up, it turns out that the reflected components all cancel out perfectly, and the total forward-traveling current wave adds up to I, $2I$, $3I$, and $4I$ at each transmission line section as shown in figure 3.3. This is all done under the assumption that the phases of each waveform at each node are combined perfectly in-phase, and that the impedances are perfectly matched. The reflection coefficients are, from left to right, -1/3, -1/5, and -1/7.

Looking back at figure 3.2, it can be seen that the lengths of each transmission line segment are symmetrically matched to the complementary segment of the same characteristic impedance on the opposite side. This is critical, as it ensures that the phase shift imposed on each current path along the input tapered transmission line is perfectly cancelled out along the output. If the transmission line with characteristic

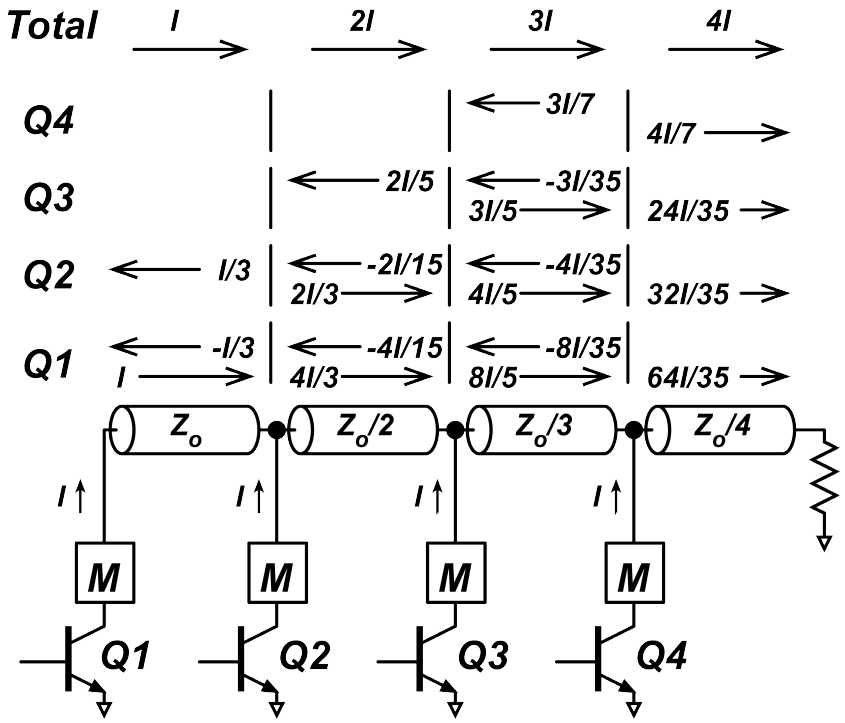

Figure 3.3: Tracking the reflected and transmitted components of the current waveforms from each transistor cell

impedance Z_o and length L_1 has a phase delay τ_1, the transmission line with impedance $\frac{Z_o}{2}$ and length L_2 has a phase delay τ_2, etc. then the overall relative phase shift applied to each individual path is simply $\tau_1+\tau_2+\tau_3$. In an ideal lossless case, it is not completely necessary to make Z_o on the input equal to Z_o on the output as long as the dielectric material in each transmission line is the same, however, in realistic cases, the phase shift and insertion loss will vary with characteristic impedance slightly, which introduces unwanted asymmetry.

The key advantage of this technique over classical corporate power combiners for mm-wave power amplifiers is that, with the design of wideband matching networks, it enables a bandwidth nearly equivalent to that of a single transistor cell with input and output matching networks. The power combiners themselves are not a critical limiting factor of bandwidth in this case. As frequency varies, the input and output tapered transmission

lines remain symmetrical. This technique can also be used to cascade multiple stages to increase power gain.

The key disadvantages are that in standard IC processes, creating extremely low-impedance transmission lines requires a lot of space, and it runs the risk of propagating multiple electromagnetic modes, creating unpredictable behavior. At high frequencies, small phase mismatches can result in largely unequal distribution of power along each pathway, causing some transistor cells to saturate before the others do. This reduces the linearity and causes premature compression of the gain curve.

3.5 Sub-Quarter Wavelength Baluns

Another amplifier topology similar to a tapered-line distributed amplifier is the sub $\frac{\lambda}{4}$ balun. In this topology, the underlying concepts regarding the use of symmetry to cancel relative phase shifts and combine power amplifier cells in-phase are the same [31] [32] [33]. Rather than combining cells in parallel and adding current waveforms in phase, the sub-$\frac{\lambda}{4}$ balun sums differential voltage waveforms in series along a 3-conductor triaxial transmission line structure. The structure is reminiscent of a Marchand balun, which is shown in figure 3.4. A Marchand balun converts signals between single-ended and differential modes by coupling a $\frac{\lambda}{2}$ transmission lines to two unconnected transmission lines. This structure differs from a Marchand balun in that, the coupled transmission lines are shorter than $\frac{\lambda}{4}$. By reducing the length of the lines, a calculated shunt tuning inductance can be applied to the output of each transistor cell as each signal is being added in series [31]. In this section, I will discuss sub-$\frac{\lambda}{4}$ balun amplifier designs, and their advantages.

The sub-$\frac{\lambda}{4}$ baluns are created using a three-conductor transmission line structure. If the 3rd conductor is assumed to be electromagnetically shielded from the 1st conductor,

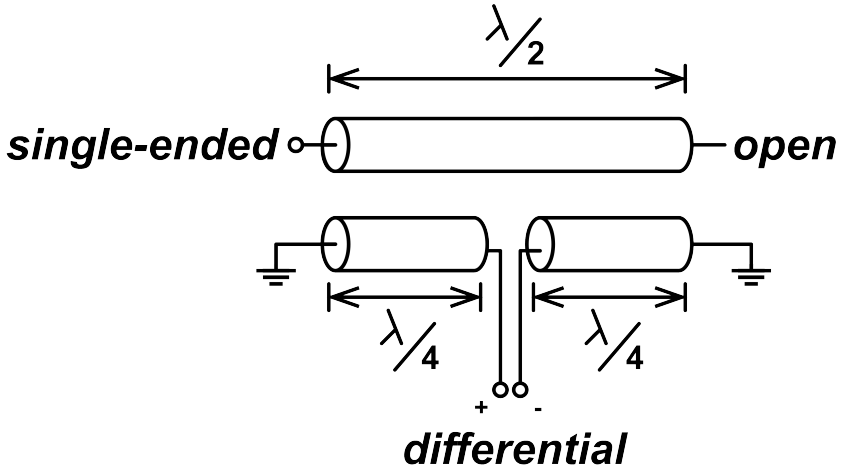

Figure 3.4: Diagram of a simple Marchand balun

then the structure will behave as two independent transmission lines [31]. Figure 3.5 shows an example 3-metal structure adding two differential voltage sources in series in the transmission line formed between m2 and m3. The lower transmission line between m1 and m2 behaves as a shunt transmission line tuning element. This element can be approximated as an inductor if it has a length less than $\frac{\lambda}{4}$. By carefully tailoring the length of this lower transmission line segment, this three-metal structure can be used to simultaneously contribute to an output impedance matching network and combine the output of a large number of transistor amplifier cells in-phase [31]. If the size of the transistor power cell and the bias conditions are chosen carefully, these structures can actually completely replace the output matching network, thereby maximizing bandwidth and minimizing output network insertion losses.

Similar to the tapered-line distributed amplifier topology, the sub-$\frac{\lambda}{4}$ balun topology requires that the upper transmission line impedance be adjusted at different points to accommodate the correct ratio of voltage and current [31]. In this case, since voltages are adding instead of currents, there is an impedance increase instead of decrease as power from each transistor cell is added. In both of these topologies, the limitation

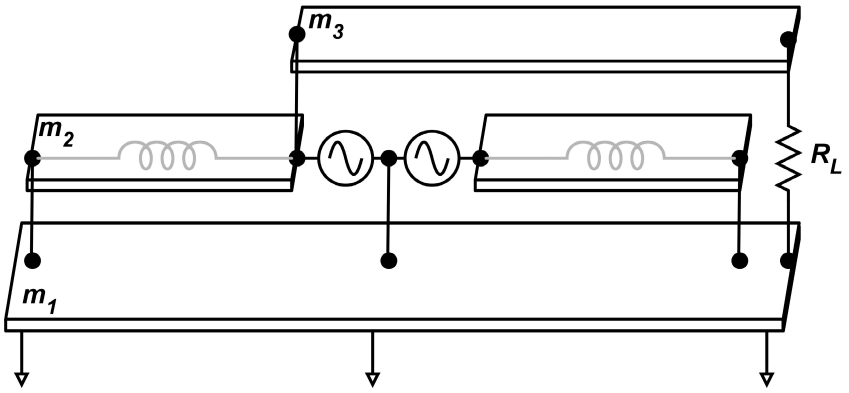

Figure 3.5: 3-metal transmission line structure

on the number of cells to be combined is that it is difficult to create spatially efficient transmission lines with very low impedances. One of the advantages of combining cells in series instead of parallel is that with series-combining, two cells can be combined in parallel with horizontal symmetry, doubling the amount of power-combining without making the transmission line impedances become unreasonably low. In the previous case where cells are being combined in parallel already, Z_o must be large so that the transmission line impedances do not become extremely small when the characteristic impedance is $\frac{Z_o}{4}$ in the case of 4-way combining. If another set of 4 transistor cells were added on the other side of the tapered transmission line combiner, the added current would be doubled, and the lowest transmission line impedance would be $\frac{Z_o}{8}$ instead. In the series-combined case, where Z_o is the output impedance for each transistor cell, the voltage waveforms add, which means the tapered transmission line segments are now Z_o, $2Z_o$, $3Z_o$, and $4Z_o$, so Z_o must be very low. If another 4 cells are combined in parallel on the opposite side of the combiner, then the output impedance from each transistor cell can be doubled, making it a more manageable value [31].

Typically, to achieve a large bandwidth in an amplifier, it is necessary to adjust the design in ways that compromise the amplifier's maximum performance. For example, one way to increase bandwidth would be to offset the center frequency of different amplifier

stages, thus creating a wider flatter passband for the overall small signal gain curve. The problem with this for a power amplifier is that the large signal bandwidth would have much better output power for the frequencies near the center frequency of the output stage, and effectively the large signal bandwidth is much lower than the small signal bandwidth because of this. Our design method for sub-$\frac{\lambda}{4}$ balun amplifiers aims to maximize large-signal bandwidth by providing the absolute minimum amount of elements on the output matching network of the output stage. Each transistor amplifier cell has an output capacitance approximately proportional to the amount of emitter length or gate width within the transistor cell and a load resistance. The simplest possible matching network would be the RLC filter in figure 3.6, where the capacitance is simply the internal device capacitance. In this case, the transistor cell must be driven at the appropriate bias conditions such that the transistor load line impedance $(V_{max} - V_{min})/I_{max}$ is equal to the load resistance. If the inductor is chosen based on the frequency and the capacitance to have zero reactance, the bandwidth of this network is

$$\Delta f_{output} = \frac{1}{2\pi R_L C_{out}} \tag{3.4}$$

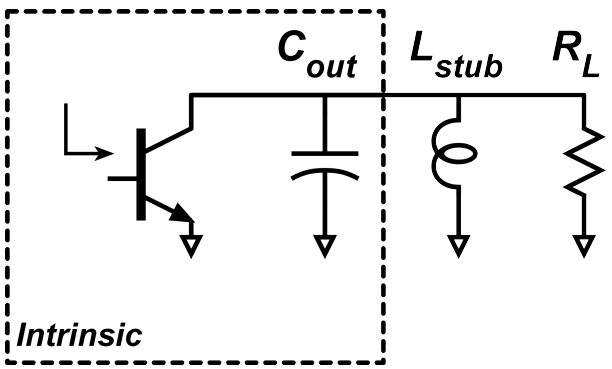

Figure 3.6: simplest, most wide-band single-stage matching network

By carefully choosing the transistor cell size and the bias conditions to match the load impedance, we can effectively use the baluns as a matching network and large power

combiner simultaneously [31]. The structure itself is more compact than a Marchand balun and also has lower insertion losses. Of course, the inductor used as a tuning element in figure 3.5 is an approximation. In reality, that inductor is a transmission line. Therefore, the bandwidth of this tuning network is also dependent on the frequency ranges for which the inductor is a valid approximation. If the bandwidth extends high enough, the transmission line will approach $\frac{\lambda}{4}$ and the approximation will deviate from reality. Otherwise, the bandwidth limit is determined by the amount of capacitance per unit emitter length or gate width inherent to the transistor technology [31].

If the method of design is to size and bias the transistors to match the load line impedance and use the balun as a matching network, then another limiting factor is the amount of series inductance which comes between the transistor cell and the balun port. This can be visualized by looking at figure 3.7. The transistor cell has some finite width and must have some space separating it from other transistor cells. Therefore, there will need to be some section of transmission line connecting the transistor output to the balun port. This results in an unwanted impedance transformation, which can be seen in figure 3.8. In this figure, there are points on the (admittance space) Smith chart numbered 1 through 4. The chart is normalized to the load impedance of each transistor power cell. The first point is the load line resistance, which is equal to the amplifier cell load impedance by design with careful choice of bias conditions. The second point represents the effect of the unwanted series inductance - a clockwise rotation around the Smith chart. In order for the tuning effect of the balun to match the output of the amplifier, the admittance must return to the normalized conductance = 1 circle. This can be done with passive elements or by simply reducing the drive current, resulting in the lateral shift shown between points 2 and 3. The admittance moves to the center of the Smith chart with a shunt inductive transformation from the sub-$\frac{\lambda}{4}$ balun. If the unwanted inductor L_{series} is larger, then the rotation between points 2 and 3 on the

Smith chart will be greater, which means a larger reduction in drive current is needed to bring the impedance back to the normalized conductance = 1 circle. A lower drive current results in a lower output power, therefore it is desirable to make this section as short as possible.

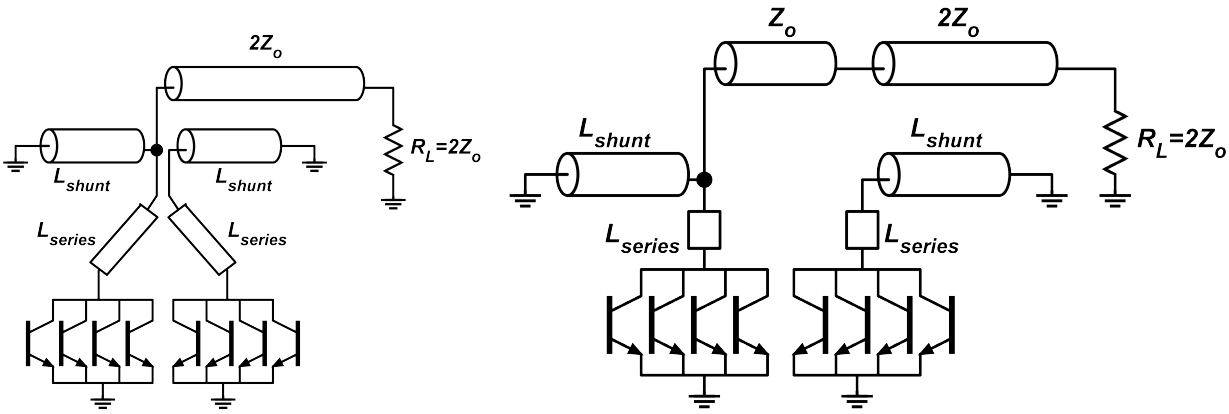

Figure 3.7: When the transistor cells are large, the geometry results in unwanted series inductance (Left). This can be corrected by adding an intermediate T-line section between ports (Right)

At low enough frequencies, the series inductance can be manageable, but at high frequencies, it can be large enough to cut severely into the drive current or even pull the admittance at point 3 into the north side of the Smith chart. In this case, this effect must be remedied. One method is to simply apply a shunt transmission line element right at the transistor cell output to match the impedance right away, and then apply a shunt capacitance right at the balun port to cancel out the balun's inductance. The advantage of this is that it is the most effective way to neutralize this series inductance and that it allows a little more freedom in the design of the balun structure. The downside is that the multiple shunt elements in this case result in larger output losses, and the bandwidth of the output tuning network will be reduced by the larger number of matching elements. Another effective strategy is to abandon the idea of differential signal paths and to apply transmission line sections with precisely tailored characteristic impedances depending on

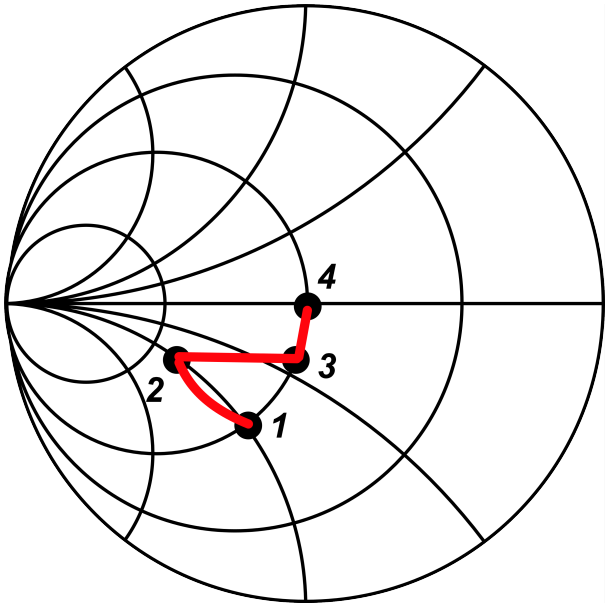

Figure 3.8: Step-by-step admittance transformations seen in design flow of sub-$\frac{\lambda}{4}$ balun amplifiers

how much current and voltage the structure is carrying at that point. This strategy results in an amplifier topology which very closely resembles a classical tapered-line distributed amplifier. A block diagram of this method can be seen in figure 3.7 on the right.

Chapter 4

Analog Beamformer Matrix IC Designs, Simulations, and Limitations

4.1 Introduction

I spent a large portion of 2016 designing a scalable analog beamforming matrix as a component of a multi-beam phased array for spatial multiplexing. The block can be implemented along with modular circuit blocks from more classical single-beam phased arrays to form a functional spatially multiplexed transmitter or receiver for mm-wave RF bands. We aim to scale this design upwards to a 16-element multi-beam phased array, but currently the beamforming matrix is designed for 4 elements. The circuit breakout was fabricated in the Global Foundries 45 nm silicon SOI CMOS technology. The strategies and details of the circuit design will be outlined in this section, as well as the advantages and limitations and some simulations.

4.2 Theory and Architecture

The proposed 4-element multi-beam phased array system is as shown in figure 4.1. As in previously described topologies, there will need to be a multi-stage RF LNA at the front of each antenna element to reduce the effects of the noise contribution from the downconversion mixer. The output from the LNAs will lead to a downconverting quadrature vector modulator and the high-frequency components are removed with a low-pass filter. This yields the inputs to the baseband beamforming matrix IC - an I and Q component for each antenna element. Recall 2.5 and 2.6 from section 2.5 regarding spatial multiplexing of multiple signals. Since we intend to use this architecture to simultaneously receive 4 signals, the linear matrix $\mathbf{M}$ tells us what the magnitude and phase is for the contribution of each individual channel is to the overall signal carried on each antenna element. To perfectly separate these channels, ideally we would apply the linear inverse matrix $\mathbf{M}^{-1}$ to the vector $\vec{I}(t)$. Therefore, our strategy is to apply a variable positive or negative gain between each input I or Q component (the I and Q components of the data from each antenna element) and each output I or Q component (the I and Q components of the data carried on each individual channel). This effectively provides a variable magnitude and phase transformation between each input antenna element and each output data channel.

The circuit for the baseband analog beamforming matrix consists of 8 individual differential inputs and 8 differential outputs. This is for 4 different I and Q input and output paths for 4 antenna elements and 4 output channels respectively. There is an array of differential folded gilbert cells, which work together to provide precise binary positive or negative gain modulation from each input I or Q pathway to each output I or Q pathway. The circuit diagram for the entire analog beamforming matrix IC can be seen in figure 4.2. As can be seen, it consists of a 4X4 array of 2X2 I/Q switching cells.

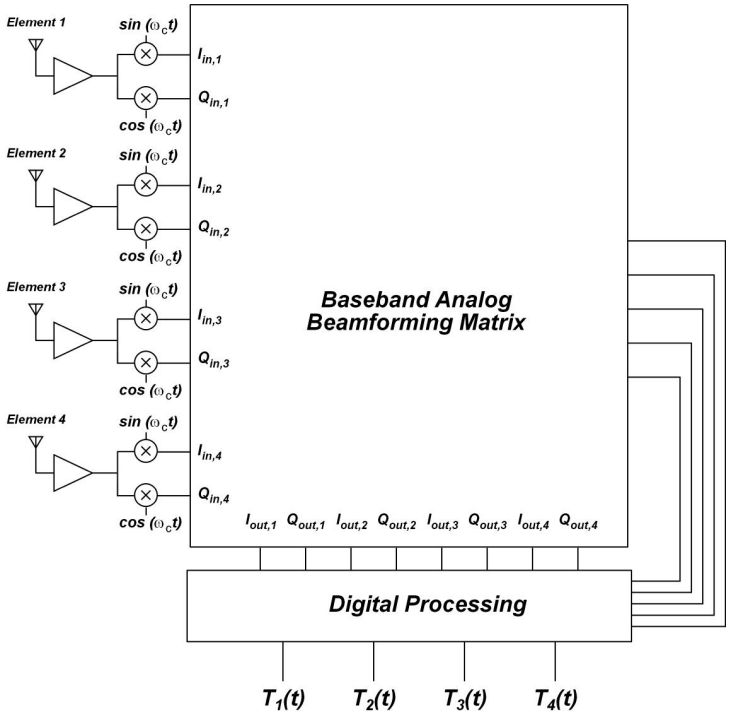

Figure 4.1: Overall system architecture for proposed low-power multi-beam phased array IC

The block diagram of each 2X2 switching cell can be seen in figure 4.3, which consists of 4 folded gilbert cell circuits. The folded gilbert cell circuit diagram can be seen in 4.4. The folded gilbert cells are intentionally designed to have a voltage gain lower than 1.

The switching cells consist of scaled folded gilbert cells, such that the analog inputs and outputs are DC coupled to the ground voltage. In most variable gain amplifiers, the DC control voltages modulate the amount of current fed to a DC biasing circuit or current mirror, thereby changing the gain. Typically, these are modeled such that the gain modulation is in nearly consistent increments. One problem with this is that for a large range of possible gain states, the current mirror transistor will not have perfectly linear behavior. In other words, it is difficult to ensure that the gain difference between adjacent gain settings will be the same between all settings. The other issue is that the transconductance variation of the switching transistors has a frequency dependence. Therefore as the operation frequency is varied, the consistency in gain variation degrades,

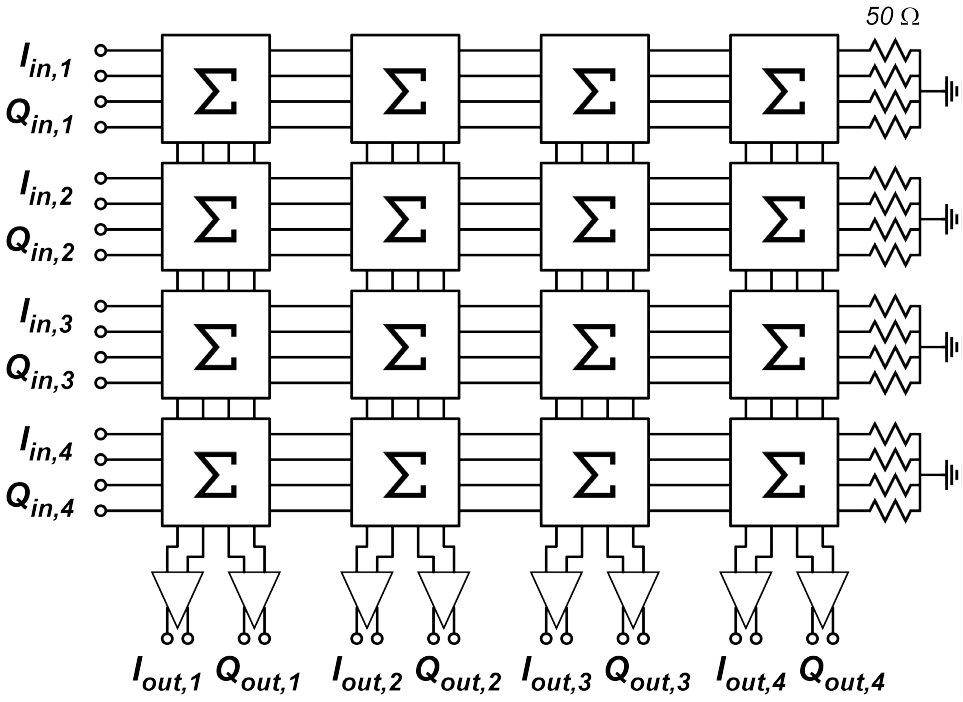

Figure 4.2: Analog beamforming matrix IC block diagram

thus limiting bandwidth. In this design, the switching is done using multiple folded gilbert cells driven in parallel. The G_m of the switching transistors are not varied by modulation of the supply currents, but rather by switching the setting of identical parallel switching cells with identical current densities and a 2^n scaling in the number of transistor fingers. This way, each cell has exactly twice the gain as the next size smaller and exactly half the gain of the next size larger, thus providing perfectly binary gain modulation. This can be seen in figure 4.9 in the next section, which shows the real and imaginary parts of Y_{21}, which is the output current divided by the input voltage vs. frequency for each of the 4 designed G_m blocks. These simulations include PEX extracted parasitic resistances and capacitances. Note that at all frequencies, both the real and imaginary part of Y_{21} scale linearly.

The trade offs in using multiple G_m blocks in parallel to modulate the gain are that the switching cells will consume more space and that they will consume more power, since

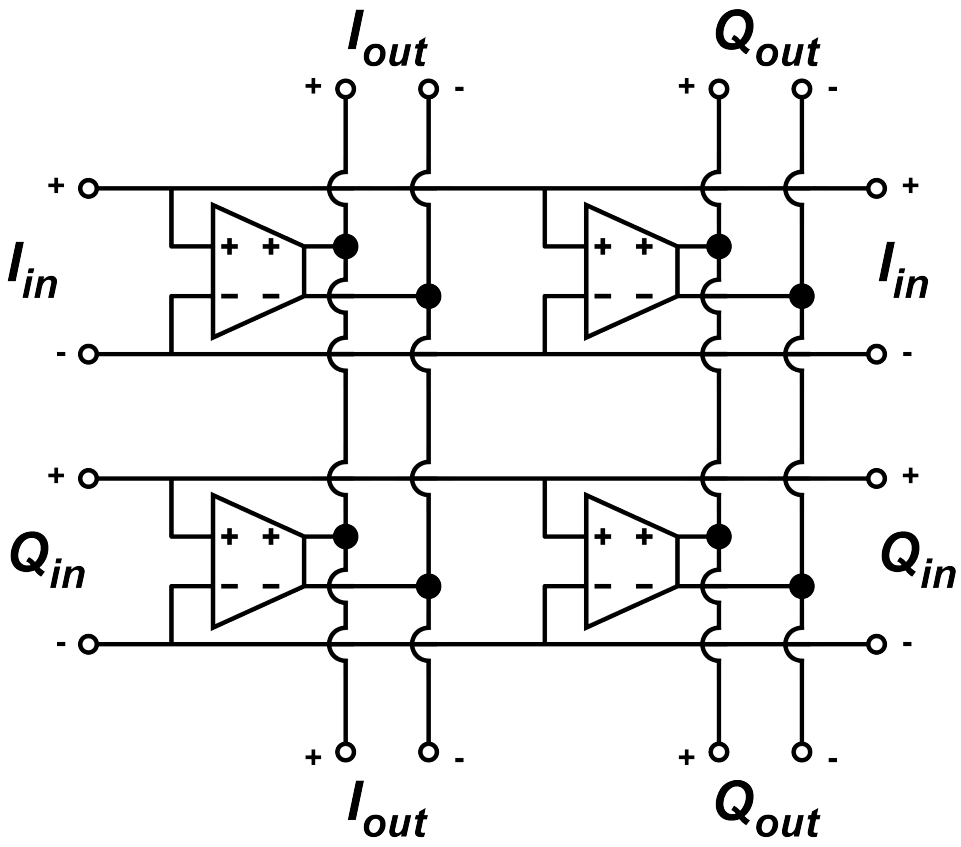

Figure 4.3: 2x2 I/Q switching cell block diagram

the power consumption will not be reduced for low gain settings. The space consumption issue, however, is negligible, since the space that each matrix cell consumes is limited by the size of the antenna elements themselves for any RF carrier frequencies below 500 GHz, assuming a $\frac{\lambda}{2}$ spacing between elements. The power consumption issue is manageable as well, but not negligible since the goal is to create a design which can be scaled upwards to as large an array as possible. The circuit operates at baseband frequency on the order of 10 GBd (10 Gsymbols/second) or lower, and the power consumption is entirely dependent on symbol rate instead of RF frequency. Furthermore, in this design, the G_m blocks are designed to have low gain and low power consumption. Since the switching cells are biased such that the input and output DC voltage is 0 V, it is helpful that

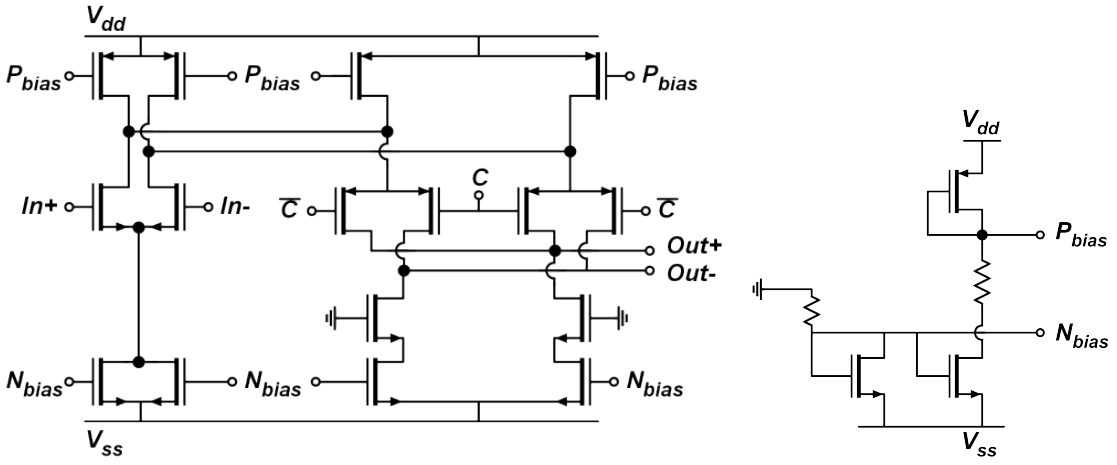

Figure 4.4: Folded gilbert cell circuit diagram (left) and DC biasing current mirror (right)

the switching cells are G_m blocks so that change in voltage due to IR drops in different gain settings is minimal. To ensure that there is very little reflection at the input, each input line is terminated with a 50Ω resistor to ground, since the RF input impedance to each switching cell is nearly infinite at low data rates. The cells generate output current, which are added together for each output I or Q path and fed to a buffer amplifier with a 50Ω input impedance and a current gain of 3. A circuit diagram of the buffer amplifier can be seen in figure 4.5. For each output buffer amplifier, there is a current mirror used for DC biasing identical to the one shown in figure 4.4.

A circuit diagram of a single folded gilbert cell can be seen in figure 4.4 on the left. The switching cells use body contacted NFETs and PFETS with a gate width of 700 nm and 875 nm respectively. The binary scaled switching cells include folded gilbert cells using 1-finger transistors, 2-finger transistors, 4-finger transistors, and 8-finger transistors, so there are 4-bits per matrix cell and 16 possible binary phase states. The current mirror used in each matrix cell is shown in figure 4.4 on the right. A serially loaded master-slave D-latch flip-flop is used as a shift register to load the digital data that controls the gain of each switching cell.

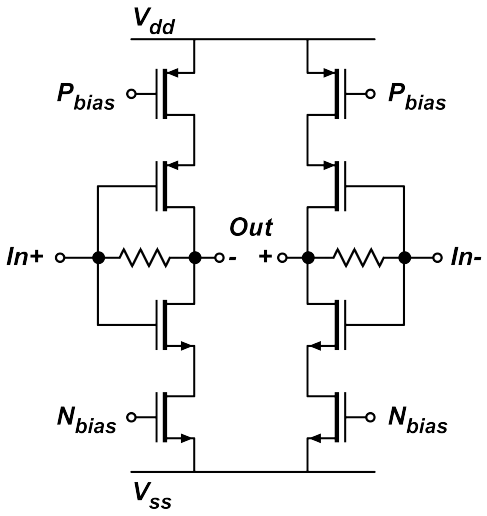

Figure 4.5: Circuit diagram of output buffer amplifier

The folded gilbert cell using single-finger body-contacted transistors has a length of $56nm$, a width of $700nm$, has a transconductance of $477.8\mu S$, and pulls $179\mu A$ of current. We decided that it would be safer to use body contacted transistors for this block to avoid the potential for floating-body effects to change the threshold voltage of the transistors. In this way, the benefits from the SOI fabrication process were not much of an advantage. However, since we ultimately intend to implement this block in a monolithic front-end for mm-wave carrier frequencies, it was important that this be done in a high-speed technology that will enable us to design RF blocks with sufficient performance. The supply voltages used are $V_{dd} = 1V$ and $V_{ss} = -1V$. Each matrix cell burns $5.7mW$ of DC power. Each buffer amplifier uses much larger transistors with 60 fingers, each with a gate width of $3\mu m$. Each buffer amplifier pulls $18.4mA$ of current, and therefore burns $36.8mW$ of DC power. The overall 4-element, 4-channel analog beamforming matrix IC with all 8 buffer amplifiers uses a total of $476.8mW$ of DC power.

4.3 Limitations

At the highest gain settings, each matrix cell has a transconductance (Y_{21}) of 5.7 mS at small signal. The transconductance becomes 90% of the small signal value (5.13 mS) with an input voltage amplitude of 155.56 mV. This same input voltage results in a 10% drop in transconductance for any gain setting and for the case of 8 parallel matrix cells driving a single output. Since there are 4 separate signals, the overall linear dynamic range would indicate that the signals could be separated reliably as long as the sum of the magnitudes of the voltage waveforms was less than 150 mV or so, and that the power difference was not so large that the weakest signal was overwhelmed by the larger signals.

The IC and architecture have some key limitations that depend on several factors. One limitation is that there is a finite number of bits controlling the gain between each input I or Q path and each output I or Q path. As explained before, in the ideal case, the channels are perfectly separated if the inverse matrix $\mathbf{M}^{-1}$ is applied to the vector of data signals carried by the input signals for each antenna element. Each element within this inverse matrix has a magnitude component, and a phase component. In other words, each element in this linear matrix is a vector which dictates the ideal case for how much gain and phase shift should be applied between each antenna element and each data channel. So insead of expressing this matrix of vectors numerically, let's instead represent it graphically with the 4X4 matrix of arbitrary vectors shown in figure 4.6.

In the ideal case, each 2X2 I/Q switching cell would provide this exact transformation (with the magnitude normalized to 1) between the appropriate I/Q input paths and I/Q output paths. In our case, we have only 4 bits of resolution determining the actual gain and phase shift which is applied. With 4 bits of resolution for each switching cell, there are a total of $2^4 * 2^4 = 256$ possible vector transformations that the 2X2 I/Q switching

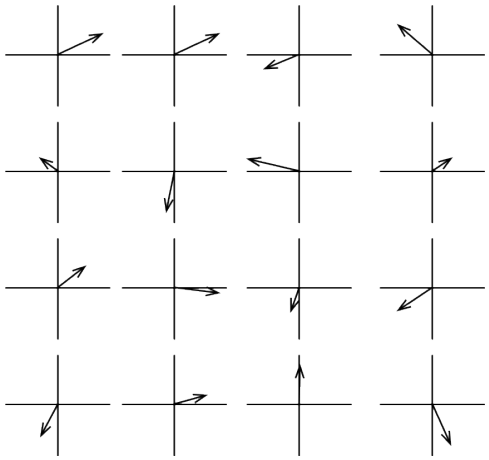

Figure 4.6: Graphical representation of transformation matrix consisting of vector elements

cell can provide. It should be noted that only analytic functions can be used as a vector transformation here, meaning the real part (I component) and the imaginary party (Q component) must are always treated the same way. In this particular case, the vector transformation from Q_{in} to I_{out} and Q_{out} is rotated 90 degrees from the vector transformation of I_{in}. Figure 4.7a shows one example of a normalized ideal transformation vector mapped onto the grid which shows the 256 possible vector transformations. Notice that the ideal transformation vector does not land exactly on a point which can actually be set with 4 bits of resolution. Instead, it has to be rounded to the nearest point. If we zoom into this square on the grid as in Figure 4.7b, we can see that there is an error vector between the actual transformation vector and the ideal transformation vector that itself has magnitude and phase.

For the case where the transmitted signals carry some data in QPSK form, each transmitted signal will have some real and imaginary component. In this case, referring back to equation 2.5, I will treat each time dependent transmitted signal, T_x, as just the raw baseband signal, as in the carrier frequency has been dropped from consideration. Each time-dependent transmitted signal, $T_x(t)$ can be expressed as $T_x(t) = I(t) + jQ(t)$ where $I(t)$ and $Q(t)$ are +/- 1. For a 4 element matrix, such as the one I have designed

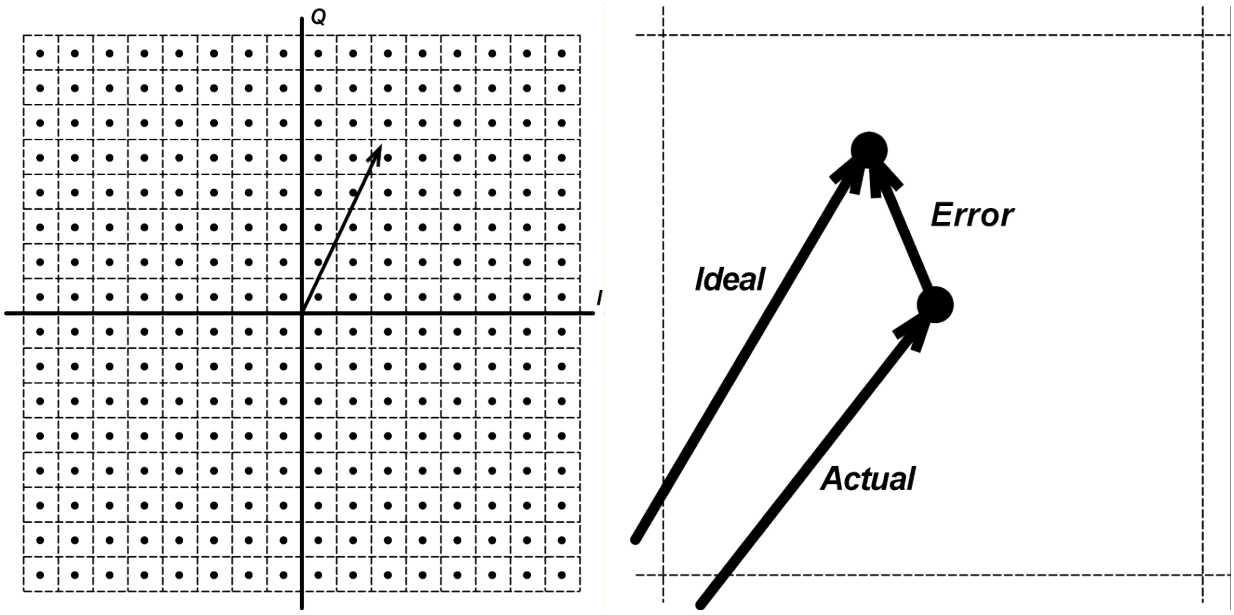

Figure 4.7: a) Ideal transformation vector mapped onto possible actual transformation coordinates with 4 bits of resolution (left) and b) zoomed in diagram of the error vector (right)

here, the vector of transmitted baseband signals can be written as it is in equation 4.1.

$$\vec{T}(t) = \begin{bmatrix} T_{R1}(t) \\ T_{I1}(t) \\ T_{R2}(t) \\ T_{I2}(t) \\ T_{R3}(t) \\ T_{I3}(t) \\ T_{R4}(t) \\ T_{I4}(t) \end{bmatrix} \tag{4.1}$$

Where each signal in this vector is an independent discrete random variable with possibilities depending on the chosen modulation scheme. Now the matrix, $\mathbf{M}^{-1}$, referred to in 2.6 which before was a 4x4 matrix of complex numbers is now an 8x8 matrix of real

numbers. Now, if we had some type of QAM modulation scheme, the time dependent signals T_{xx} would be multilevel signals, and the analysis and would each have a mean and variance. The analysis is then complicated. We consider for brevity the simpler case of QPSK modulation. Similarly, analysis for a non-orthogonal matrix in the general case would also be complicated. Therefore, we will consider the specific case where the relative phase shift between array elements for the four incident beams are 0 degrees, 90 degrees, 180 degrees, and 270 degrees. This corresponds to beams separated by the angular resolution of the array, i.e. the Rayleigh criterion. In this case, the original complex 4x4 transformation matrix $\mathbf{M}$ from equation 2.5 could be written as it is in equation 4.3.

$$\mathbf{M} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -j & -1 & j \\ 1 & -1 & 1 & -1 \\ 1 & j & -1 & -j \end{bmatrix} \tag{4.2}$$

The inverse matrix of this would be as it is in equation 4.4.

$$\mathbf{M}^{-1} = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & j & -1 & -j \\ 1 & -1 & 1 & -1 \\ 1 & -j & -1 & j \end{bmatrix} \tag{4.3}$$

which can be normalized by multiplying the entire inverse matrix by 4. Now, the received signals are exactly the same as what was defined in equation 2.5, however, due to the quantization error caused by the finite gain resolution of the G_m blocks there is inevitably going to be some amount of error in the gain settings applied to the received signals as shown in figure 4.7. Breaking up the individual data streams into the I and Q

components (real and imaginary components of $\vec{T}(t)$, the transmitted signals which are actually recovered, $\vec{T_{rec}}$ as as expressed in equation 4.4.

$$\begin{bmatrix} T_{rec,R1}(t) \\ T_{rec,I1}(t) \\ T_{rec,R2}(t) \\ T_{rec,I2}(t) \\ T_{rec,R3}(t) \\ T_{rec,I3}(t) \\ T_{rec,R4}(t) \\ T_{rec,I4}(t) \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & -1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 & 1 & 0 & -1 \\ 1 & 0 & 0 & -1 & -1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & -1 & -1 & 0 \end{bmatrix} + \epsilon_{\mathbf{g}} \right) \begin{bmatrix} R_{R1}(t) \\ R_{I1}(t) \\ R_{R2}(t) \\ R_{I2}(t) \\ R_{R3}(t) \\ R_{I3}(t) \\ R_{R4}(t) \\ R_{I4}(t) \end{bmatrix} \tag{4.4}$$

Where $\epsilon_{\mathbf{g}}$ is the 8x8 error matrix representing the error in the applied gain caused by the finite gain resolution of the G_m blocks. If we normalize the gain settings such that the edges of the box on the left side of figure 4.7 correspond to a gain of -1 or 1, then the edges of the box on the right side of figure 4.7 will correspond to $-\frac{1}{2^n}$ and $\frac{1}{2^n}$ where n is the number of bits of gain resolution in each G_m block. We can treat the error in the x-direction (real) and the y-direction (imaginary) as independent uniformly distributed random variables between $-\frac{1}{2^n}$ and $\frac{1}{2^n}$. For a uniform random variable with boundaries b and a, the variance is $\frac{(b-a)^2}{12}$. Therefore, each value in the error matrix ϵ_g has a variance of $\frac{1}{3}\frac{1}{2^{2n}}$.

For the specific case that we are considering, the time dependent received signals at each antenna element, $R_{xx}(t)$, are the sum of 4 numbers that can each be 1 or -1 at any point in time. Therefore, each received signal at each antenna element can be treated as a discrete random variable with a 1/16 chance of being -4, a 1/4 chance of being -2, a 3/8 chance of being 0, a 1/4 chance of being 2, and a 1/16 chance of being 4. The variance

of this discrete random variable is $\frac{1}{8} * 16 + \frac{1}{2} * 4 = 4$. Because the random variables $\epsilon_{g,xx}$ and $R_{xx}(t)$ are completely independent and both have a mean value of 0, we can define a variable, ϵ_{total} which is the total error in each recovered transmitted signal that has a variance equal to the product of the variance of $\epsilon_{g,xx}$ and $R_{xx}(t)$. The variance of ϵ_{total} is equal to $\frac{4}{3*2^{2n}}$. The expected value of the error in each recovered transmitted signal is therefore the standard deviation of ϵ_{total}, which is equal to $\frac{2}{\sqrt{3}*2^{n-1}}$. By converting this value to dB for 4 bits and 6 bits, it can be seen that the expected amount of isolation from other channels for each recovered signal is -22.83 dB (7.2%) for 4 bits and -34.87 dB (1.8%) for 6 bits of resolution, assuming all channels were received with equal power.

To demonstrate this, I designed a simulation testbench where 8 separate random bit streams are generated and applied to the 4-bit beamforming matrix exactly as they would if 4 orthogonal QPSK signals were received simultaneously and the unwanted frequency elements were perfectly removed. I assumed that the incident beam angles corresponded to a relative phase shift of 0 degrees, 90 degrees, 180 degrees, and 270 degrees as in the case above. I made a separate simulation block that uses 6 bits of resolution by adding a second copy of the smallest folded gilbert cell and applying an analog voltage to the control bit. I generated an eye pattern for each case. Based on the calculation above, it would be expected that the eye has approximately 1.8% closure for the 6-bit case and 7.2% closure for the 4-bit case. The simulation was run assuming a 5 Gsymbol/second bit rate. Those eye patterns can be seen in figure 4.8. The closure in this case was 2.05% for the 6-bit case and 9.02% for the 4-bit case. This is very close to the expectation, and therefore supports the calculation.

In the more general case, the variance of the received signal for each antenna element is directly proportional to the number of signals. So for a 5 element receiver that is receiving 5 signals, the variance of ϵ_{total} is equal to $\frac{5}{3*2^{2n}}$ and the expected value of the error would be $\frac{\sqrt{5}}{\sqrt{3}*2^{n-1}}$. For a 6-element receiver receiving 6 QPSK signals, the expected

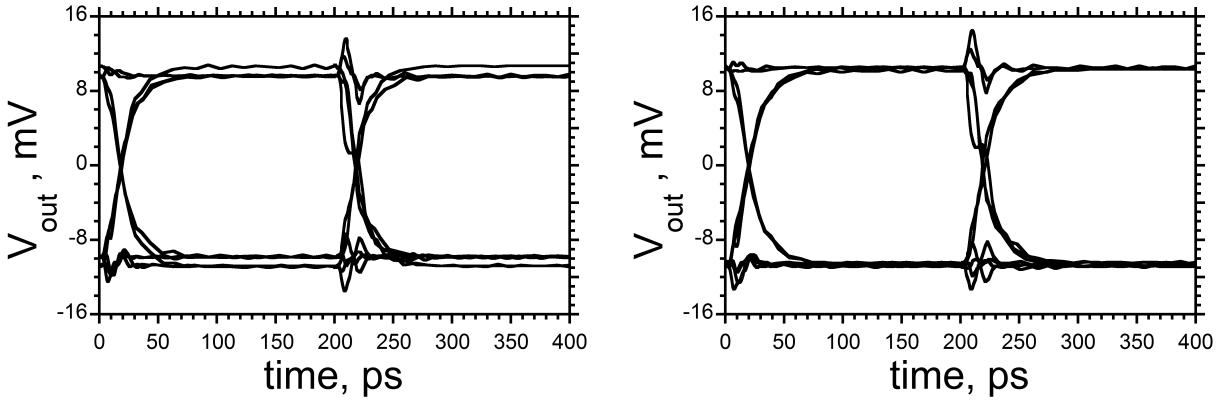

Figure 4.8: Eye patterns with ideal matrix inputs for a) 4 bits of gain resolution and b) 6-bits of gain resolution

value of error would be $\frac{\sqrt{5}}{\sqrt{3}*2^{n-1}}$ and so on. Furthermore, the variance of ϵ_{total} varies with the modulation scheme chosen. For 16QAM, the variance is 5 times larger than it is for QPSK. For 64QAM, the variance is 21 times larger. Therefore, the number of bits of gain resolution required to achieve a certain level of isolation between channels will depend on the number of antenna elements (and signals), the modulation format, and the chosen symbol-rate. In order to scale a design up to a larger number of antennas and signals and to use a modulation format with more bits per symbol, it will be necessary to use a larger number of bits of gain resolution in the switching cells of the beamforming matrix. Of course, this is not the only source of error and crosstalk between channels that need to be considered. Some frequency dependent causes will be discussed below.

The gain resolution of the switching cells can be increased in two ways. One way would be to increase the number of scaled G_m blocks, but adding more G_m blocks with larger and larger numbers of transistor fingers would unfortunately dramatically increase the power consumption and size of these circuits. A more efficient approach would be to add a second single-finger G_m block and control its gain with an analog voltage. This would cap the power consumption and allow an arbitrarily high resolution to the gain blocks,

depending on the precision of the voltage control. On the downside, a single-finger G_m block with an analog control bit in this architecture will not display the same frequency response at each gain setting. This can be seen in the real and imaginary parts of Y21 for a single-finger block controlled in this manner at several settings in figure 4.9. While this could be managed by simply mapping out the appropriate voltage control settings at the desired bit-rate, the issue still remains that the imaginary component of Y_{21} also does not scale linearly with gain. This would result in unintended and uncontrolled phase shifts from the matrix cells which are designed to provide a very precise gain and phase shift to the incoming signal. The uncontrolled phase shifts would be negligble at low data rates, but would become more problematic as the data rate is increased.

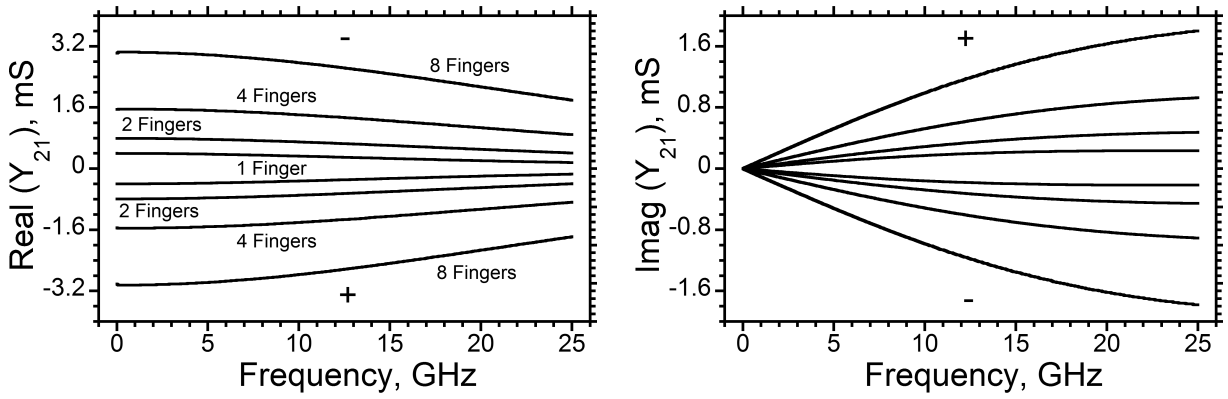

Figure 4.9: a) Real part of Y_{21} for G_m blocks of 8, 4, 2, and 1 finger with positive and negative gain setting applied (left) and b) Imaginary part of Y_{21} (right)

Another limitation on data rate comes from the passive frequency-dependent components of the circuit. While the wires between each matrix cell are designed to be transmission lines with Z_o nearly equal to 50Ω, these wires are inevitably loaded with unintended frequency dependent components which can change the input and load impedances seen by each G_m block. In a 4-element matrix like the one I've designed, these unintended impedance transformations are manageable, but in order to acheive a functional matrix

with many elements and a high data rate, the matrix will need to be designed in an extremely compact package with very precisely controlled impedances. Figure 4.10 shows the simulated real and imaginary parts of S_{21} for the highest and lowest gain settings for the lower left corner of the beamforming matrix and the upper right corner of the matrix, which represents the largest differences caused by this effect in this design. In this simulation, the reactive components are distributed throughout the circuit in the form of hybrid-pi transmission models made from lumped inductors and capacitors to model the behavior of signals on my designed IC. At 14.6 GHz, the magnitude of S_{21} of the matrix cell with the longest path is 3 dB lower than that of the matrix cell with the shortest path, and at 8.3 GHz, the difference is 1 dB. This disparity is caused purely by impedance transformations along the wires propagating the signal to the G_m blocks and by layout-level parasitic capacitances and resistances. This is a design challenge that will be extremely formidable for more highly-scaled designs. Errors caused by finite gain resolution, as explained above are not dependent on symbol rate. Errors associated with uncontrolled passive impedance transformations are. As these are the 2 major contributors to error, it is expected that lower frequencies, the quantization error caused by finite gain resolution will be the dominant source of channel interference, whereas at higher data rates, the uncontrolled frequency dependences will be the dominant source of error.

In order to precisely quantify the limitations on data rate, I've simulated this matrix IC receiving 4 QPSK signals using ideal mixers, filters, and transmission line elements to emulate 4 channels simultaneously being received. The RF frequency in this case is 60 GHz. The antenna spacing was chosen to be 2.5 mm so that it is equal to $\frac{\lambda}{2}$, satisfying the Bragg condition. The incident beam angles were chosen to be 0 degrees, 90 degrees, 30 degrees, and -30 degrees so that the relative phase shifts of each beam were 0 degrees, 180 degrees, 90 degrees, and 270 degrees. As the relative phase shifts between antenna elements are brought closer together, two things happen. The overall

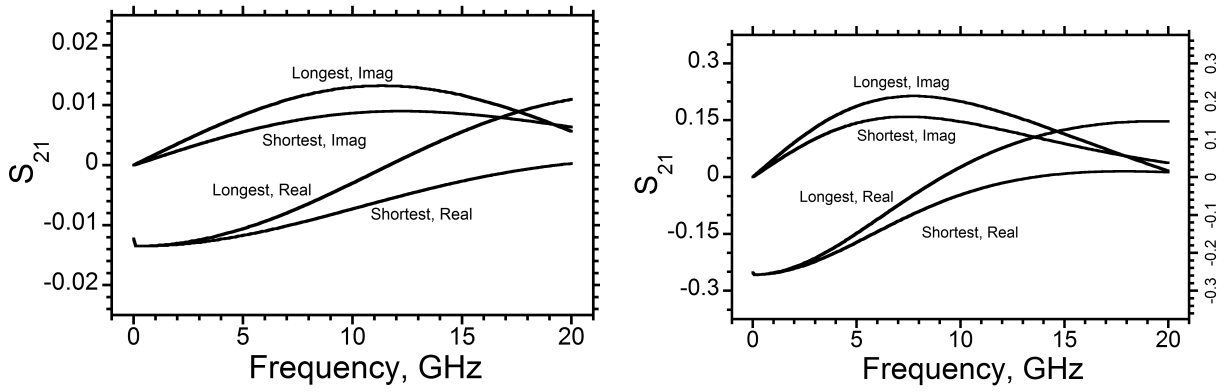

Figure 4.10: a) Real and imaginary S_{21} for the shortest path length and the longest path length with the lowest gain setting of 1 (left) and the highest gain setting of 15 (right)

data signals received by each antenna element become more similar to each other, and the magnitude component of the vectors in the inverse transformation matrix become larger and vary to a larger degree within each column. As the data signals for each antenna become more similar to eachother, it becomes harder to fully separate the independent data channels. As the magnitudes of the transformation vectors vary more, it means the magnitudes of the gain settings for some vectors will become very small after the values are normalized to 1. This means that the error vectors for these vector transformations will become larger compared to the actual gain settings, which results in even more crosstalk between channels. Therefore, to keep these simulations clean and controlled, I've used the largest possible spacing of relative phase shifts for the 4 incoming signals.

In this testbench, there are 8 separate random bit strings which are generated with +10 mV and -10 mV for 1's and 0's respectively that are applied through a 9th-order Bessel-Thomson filter. The 3-dB bandwidth of the filter is set to be 0.7 times the bit rate. The I-bit streams are mixed with a 60 GHz sine wave with no phase shift and the Q-bit streams are mixed with a 60 GHz sine wave with a 90 degree phase shift. The I and Q paths for each RF signal are added together with an ideal summing block to

create QPSK RF signals. These signals are each fed into 4 different ideal transmission line elements with delays emulating the time delays seen by received signals at each antenna element. The signals for each antenna element are then added together with ideal summing elements, then split and mixed with a sine wave and a cosine wave at 60 GHz to get the I and Q components for each antenna element. These signals are fed through a 9th-order Tchebyscheff LC low-pass filter to eliminate the high-frequency components. The gain settings for the matrix IC were calculated using a simple MATLAB code which calculates the ideal inverse transformation matrix in the form of simple complex numbers, normalizes the magnitudes of each column so that the maximum real or imaginary gain is 1, and then approximates this to the closest possible gain setting based on the gain resolution. The output of the matrix was fed into 50 ohm resistors. The output I and Q values for each channel should match the 8 data streams that were created as inputs. Figure 4.11 shows the ideal 5 GB/sec I1 bit stream input divided by 2 to offset losses in the matrix IC, plotted along with the I1 output, offset by 110 picoseconds to account for the passive delays from the testbench. It can be seen that the output clearly corresponds to the input. Figure 4.12 shows the simulated eye pattern for the I1 output bit stream with and without random bit streams applied to the other paths. Figure 4.13 shows the eye patterns at other frequencies. It can be seen here that the eye has about 50% closure 15 GHz.

Another challenge for designs of this type is to design circuitry that can handle signals of different power levels simultaneously. This type of system would likely be used as a short-range base station for high-speed mm-wave line-of-sight wireless networks. As explained in the first chapter of this dissertation, the free-space path losses for propagating waves is proportional to distance squared. Therefore, even with a small operating radius of about 100 m, a user that is only 10m away from the base station would be received by the base station with a signal 100 times more powerful than a signal sent by a user 100 m

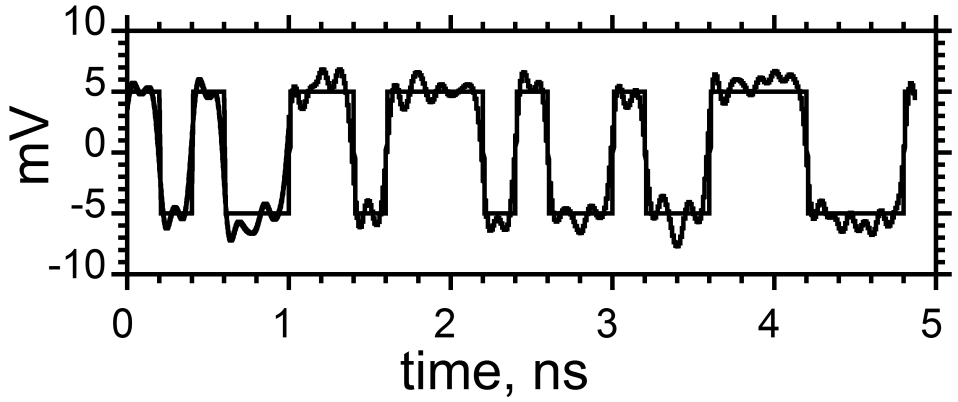

Figure 4.11: Transient simulation comparing output of I1 bitstream with the input of I1 bitstream at data rate of 5 GSymbols/sec

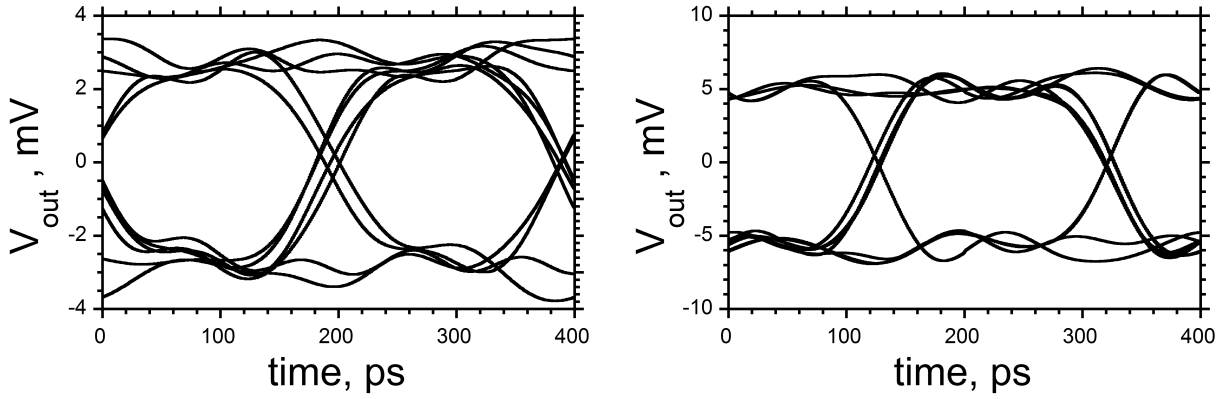

Figure 4.12: Eye pattern for I1 at 5 GSymbol/sec data rate with 8 simultaneous random bit streams (left) and with just a single random bit stream (right)

away. Unless the circuitry has nearly infinite angular resolution and gain resolution, this is a problem that will never be solved entirely by this type of hardware. This problem will, to a large degree, need to be controlled with adaptive feedback control. This means that the base station will need to sense that the power level of the signal from 10 m away is 20 dB more powerful than the signal from 100 m away and send the handset a signal which tells the handset to reduce it's transmit power. In order to figure out how much precision was needed for the adaptive component of the system, I simulated the eye

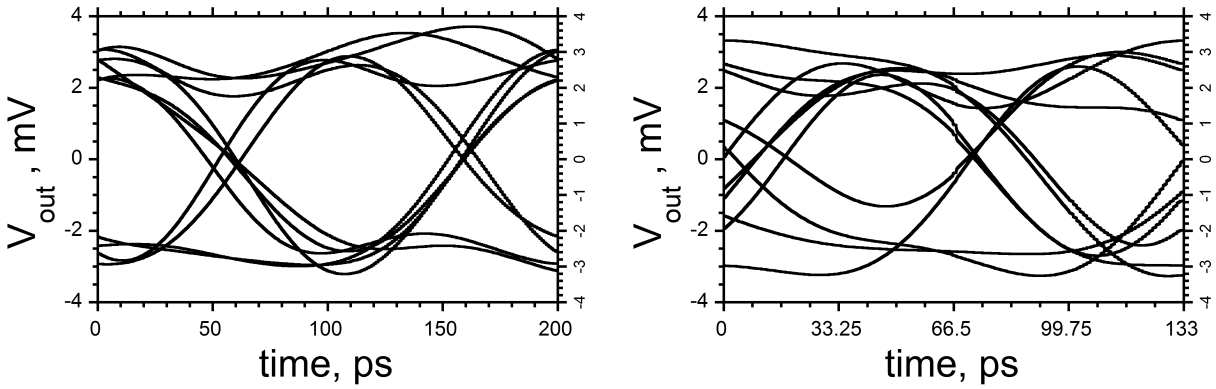

Figure 4.13: Eye pattern for I1 at 10 GSymbol/sec data rate with 8 simultaneous random bit streams (left) and 15 GSymbols/sec (right)

pattern for the case where the tracked data is 12 dB weaker than the other channels, and the case where the tracked data is 18 dB weaker than the other channels at a data rate of 5 GSymbols/second. The result can be seen in figure 4.14. It can be seen that the eye has about 25% closure when the measured signal is 12 dB weaker than the other signals and that it has about 50% closure when the signal is 18 dB weaker than the other signals at 5 Gsymbols/second. Of course, at higher data rates and with modulation schemes with more bits per symbol, the adaptive feedback control will need higher resolution to account for the reduced isolation.

There are multiple measurement plans in place to determine the maximum gain of each switching cell at different baud rates, to determine the crosstalk between adjacent channels, to determine the dynamic range of each pathway, and to demonstrate functionality. Each individual path can be measured with on-wafer GSGSG probes to determine the phase shift and gain for each differential input and output. A full channel (both an I and a Q pathway) can be measured using a single-layer test-board consisting of gold on AlN and wirebonding the IC on. The board can be probed directly with microwave "GSGSGSGSG" probes to test the gain from any input path to any output path at

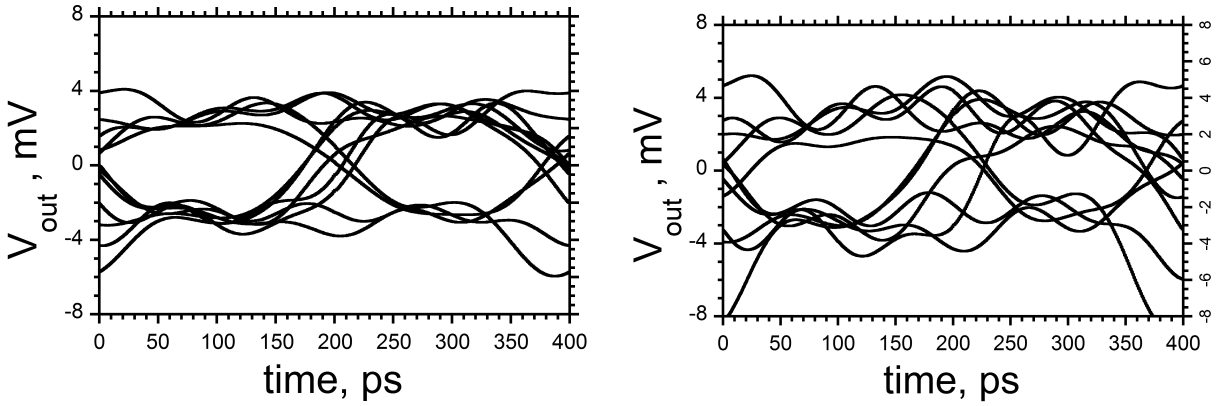

Figure 4.14: Eye pattern for I1 at 5 GSymbol/sec data rate, 3 dB weaker than other 3 signals (left) and 6 dB weaker than other signals (right)

various data rates and to demonstrate phase shifting. This single-layer test board can be wire-bonded onto a larger PCB apparatus with SMA connectors on the perimeter in order to perform multi-channel measurements, quantifying crosstalk between adjacent channels. Unfortunately, the large wire-bonding inductances will potentially limit the data rate used in measuring the breakout.

Chapter 5

DARPA ACT Designs

5.1 DARPA ACT Overview

In this chapter, I will go into detail about high-linearity and power amplifiers that I have designed as a part of the DARPA ACT program. The goal of this project was to design components for a high-dynamic range dual-conversion receiver with a broad 1-50 GHz RF tuning range [34]. The block diagram of the receiver can be seen in figure 5.1. The amplifiers that will be discussed in this thesis are highlighted in red. We wanted to achieve an extremely wide potential RF bandwidth by upconverting to a 100 GHz 1st IF frequency and driving the upconversion mixer with an ultra-wideband power amplifier. There is a common module and an application-specific LNA which can be more narrow-band and centered at the RF frequency-of-interest. By having an all-purpose transceiver module with a wide potential RF bandwidth, the design time and cost can be cut down significantly, as only one amplifier needs to be designed. This is intended to be used in military applications, so it is critical that the IP3 be as large as possible to resist distortion generated from jammers lying in the receiver's passband. The bandwidth of the system is determined by the bandwidth of the LO driver amplifier. To achieve a

high dynamic range, it is critical to achieve high linearity, particularly in the passive upconversion mixer and the 1st gain stage of the IF chain. To get a large linearity from the upconversion mixer, the wideband power amplifier driving the local oscillator port must produce a large power output over the entire bandwidth.

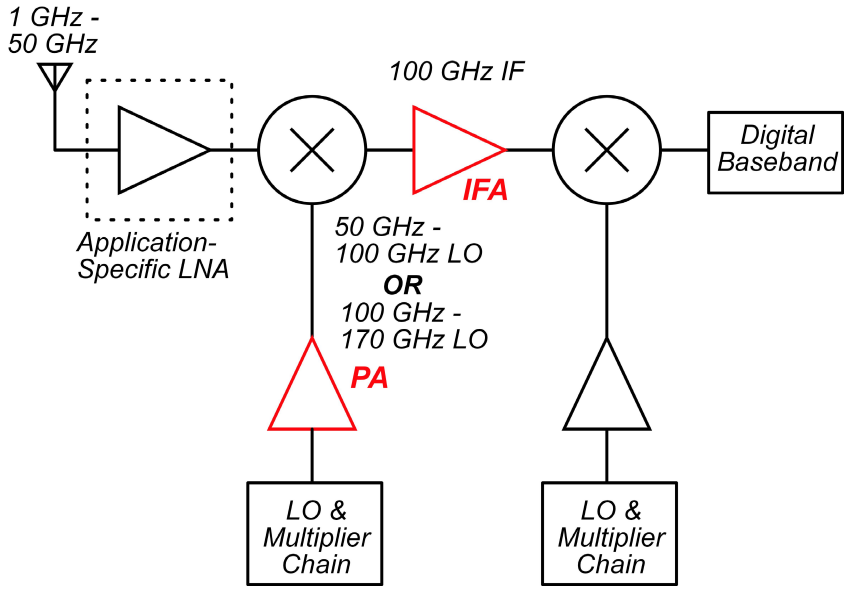

Figure 5.1: Block diagram of DARPA ACT Receiver

There are two possible frequency plans - one for low-side injection and one for high-side injection. The injection refers to the frequencies which are used for the local oscillator port of the upconversion mixer. To achieve as wide an RF tuning range as possible, the LO driver bandwidth needs to extend from 100 GHz down to as low a frequency as possible for a low-side injection design or from 100 GHz up to as large a frequency as possible for the high-side injection design. The advantage of the low-side injection plan is that the driver amplifier is at a lower frequency, meaning it is easier to get a larger output power and efficiency. The advantages of the high-side injection plan is that the driver amplifier can achieve a wider bandwidth and this frequency plan will produce less spurious tones that can interfere with the IF passband.

The amplifier ICs described here were simulated using Advanced Design System (ADS) and an HBT model provided by Teledyne. All transmission line structures, baluns, MIM capacitors, device feed structures and probe pads were simulated using ADS Momentum 2.5-D electromagnetic simulator.

5.2 TSC 130nm InP HBT Process

These designs were all designed in Teledyne Scientific Company's 130nm InP HBT process. The HBTs have a 3.5 V breakdown voltage and an f_t/f_{max} of $520GHz/1.1THz$. The process uses a 3-metal layer gold interconnect. Microstrip ground planes are in metal 1 and lines are in metal 3. The inter-metal interconnect separation dielectric material is BCB ($\epsilon_r = 2.7$). The separation between the first and second level interconnect is 1.0 μm and the separation between the second and third level interconnect is 5.0 μm. MIM capacitors ($SiN_x 0.3fF/\mu m^2$) are formed between the first and second level interconnect metal. Each metal layer is 1.0 μm thick, giving the standard thin-film microstrip transmission line 7.0 μm of separation between signal plane and ground plane. $50-\Omega/square$ thin film resistors are also available.

5.3 High-Linearity Amplifier 1st Design

A microscope photo of a breakout structure for the first design for this amplifier is shown in figure 5.2 on the left. The full IC size is 1.1 mm x 0.72 mm, while the core differential amplifier is 0.46mm x 0.47 mm. The design goals for this amplifier were to achieve an extremely high IIP3 ($> 24dBm$), a gain larger than 6 dB, and as low a noise figure as possible. Secondary goals were to have low power and space consumption. The common-emitter amplifiers were designed in a pseudo-differential topology, as this mini-

mizes interstage coupling through the power supply. Since the receiver has multiple gain stages, this is critical. Further, in a pseudo-differential structure, the impedance of the on-wafer MIM power supply bypass capacitors does not de-tune the output impedance-matching network.

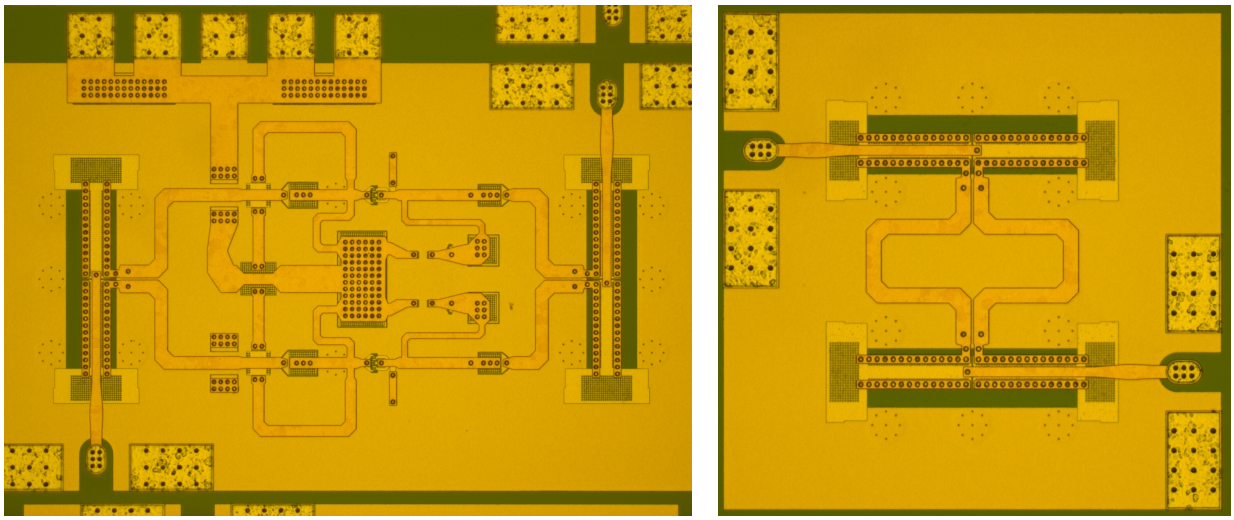

Figure 5.2: Microscope photo of amplifier breakout IC (left) and Balun test-structure used to de-embed amplifier performance (right)

A microstrip transmission line is used to terminate the emitter port of the cell to add inductive degeneration. At the cost of reduced gain, this increases the OIP3 and permits simultaneous matching for low noise and low input reflection coefficient. At the collector port of the cell, a $\frac{\lambda}{4}$ transmission line short-circuits any locally-generated 200 GHz second harmonic, as this mixes with the fundamental to produce additional 3rd-order distortion. The base port of each HBT cell is biased using a simple current mirror. A circuit block diagram can be seen in figure 5.3. The HBT cells in the amplifiers were composed of 4 emitter fingers, each with $L_e = 5\mu m$. The emitter fingers within the cell are spaced at a large 8.7 μm for reduced thermal resistance. The quiescent HBT bias was $I_c = 100mA$($2mA/\mu m$ of L_e) and $V_{ce} = 2V$.

Input and output baluns were added to permit on-wafer testing of single-stage ampli-

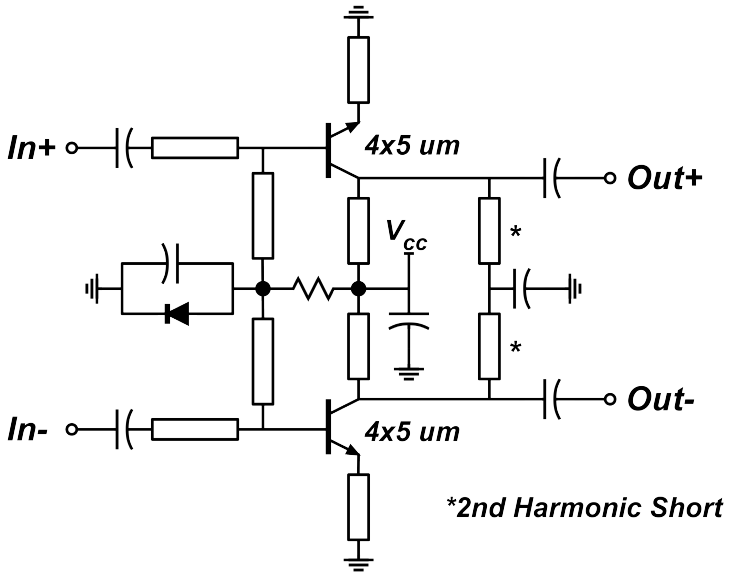

Figure 5.3: Circuit diagram of 100 GHz high-linearity amplifier

fier test structures. The baluns are sub-$\frac{\lambda}{4}$ with a shunt matching capacitor to bring them into resonance at 100 GHz. The insertion loss of the baluns and corresponding matching capacitors were measured to be 1.2 dB using a separate test structure which can be seen in figure 5.2. The measured insertion losses of the baluns were used to de-embed the gain, noise, and IP3 of the core differential amplifier. Both the breakout performance and the de-embedded amplifier performance will be presented in the following section.

5.4 High-Linearity Amplifier 2nd Design

A new high-linearity amplifier design was designed for a center frequency of 94 GHz. The IC was designed using much of the same strategies as the previous attempt. The key difference is that a larger transistor cell was used ($8x5\mu m$ instead of $4x5\mu m$), and a lower current density was used to reduce power consumption. In addition to using a second harmonic short termination at the output of the transistor cell, another one was added at the input of the transistor cell as well. 2-tone harmonic balance simulations show a

linearity improvement of 0.5 dB without any performance drawbacks. The IC breakout layout design, shown in figure 5.4 as a screen capture from ADS, is $1.42mmX0.8mm$ in area including baluns and pads. Another improvement in the design is that the base biasing network now has independent control from the collector bias voltage, allowing for more freedom in biasing for measurements. A parallel RC filter was used at the input of the amplifier to improve out-of-band stability at 30 GHz.

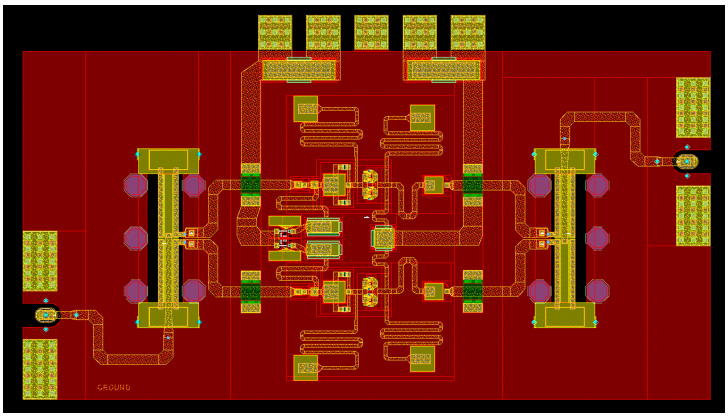

Figure 5.4: Screen capture of 2nd high-linearity W-band amplifier design

The new design simulations show 123 mW of DC power consumption, 6.7 dB of small signal gain at 94 GHz, and an OIP3 of 34 dBm (IIP3 of 27.3 dBm). The simulations show a noise figure of 6dB, an S_{11} below -15 dB, and an S_{22} of -6.6 dB. Since the output is impedance-matched for maximum linearity, an output return loss above -10 dB is expected. Overall, this new design attempt should demonstrate a marked improvement over the previous design attempt in gain, noise performance, linearity, and power consumption. The IC was submitted for tapeout in early February, and should be returned in May for measurement.

5.5 Ultra-wideband Power Amplifier Lowside Injection Design

The HBT power cells used in the amplifiers were composed of 8-emitter fingers, each with $L_e = 5\mu m$. The footprint for the power cell is $70\mu mX15\mu m$ and arranged in a common-emitter configuration. The transistors were spaced conservatively because there was no thermal model used in simulation. The quiescent HBT bias used in the amplifiers was $I_c = 5.5mA/\mu m^2$ and $V_{ce} = 2V$.

A microscope photo of a breakout structure for the low-injection power amplifier design is shown in figure 5.5. The amplifier uses basic common-emitter configuration for each HBT cell. It uses sub-quarter wavelength balun pairs for compact, wide-band power combining, for inductive load-line matching at 85 GHz, and for DC biasing. The output stage uses 4:1 power combining for maximum output power, and the driver stage uses 2:1 power combining to boost overall gain and provide additional power to the output stage. Output impedances of the output baluns are 50 ohms. The baluns present to the transistor cells a 25 ohm load, with an inductive shunt tuning of the transistor capacitance. A more detailed description of this topology can be found in [31]. A circuit diagram of the power amplifier can be seen in figure 5.6.

The large bandwidth under small signal operation is a consequence of both the high transistor bandwidth and the specific power-combiner used. In typical corporate-combined power amplifiers, the combining networks limit the bandwidth to well below the limits inherent within the transistor. In contrast, with series-combined designs presented here, each transistor is loaded with 25 ohms, in parallel with a shunt tuning inductance selected to resonate the transistor output capacitance at the design frequency. The tran-

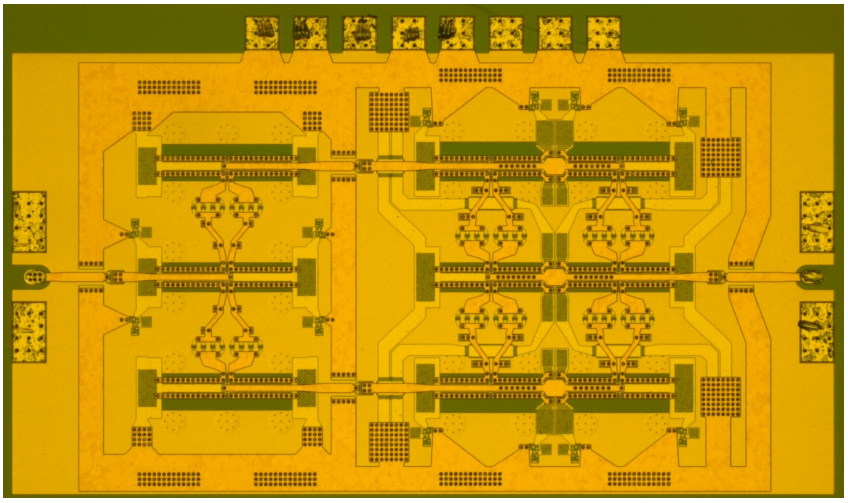

Figure 5.5: Microscope photo of 2-stage low-side injection LO driver breakout

sistor load impedance is selected so that

$$Z_{load} = \frac{(V_{max} - V_{min})}{I_{max}} = 25\Omega \tag{5.1}$$

where V_{max} and V_{min} are set close to the transistor breakdown and minimum (kirk-effect) voltages, and I_{max} to the maximum safe transistor currents. The transistor output capacitance is approximately $C_{cb} = 0.8fF/\mu m$ of L_e. Hence, from equation 5.1, using values $C_{cb} = 0.82fF/\mu mL_e$, $I_{max} = 2mA/\mu m$, $V_{max} = 3.5V$, and $V_{min} = 0.5V$, we find that $\Delta f_{output} = 130GHz$. Noting that the sharp 30 GHz low-frequency cutoff arises from a resonance between the balun inductance and the collector power-supply bypass capacitance, the measured amplifier bandwidth is close to that estimated for the output tuning network.

The amplifier bandwidth is also limited by that of the input tuning network (fig 5.6). Although the maximum feasible input tuning bandwidth is $\Delta f_{input} = 1/2\pi R_{bb}C_{be}$ which is larger than 200 GHz in this technology, the input matching network employed uses only three LC sections, and its bandwidth therefore falls well below this limit.

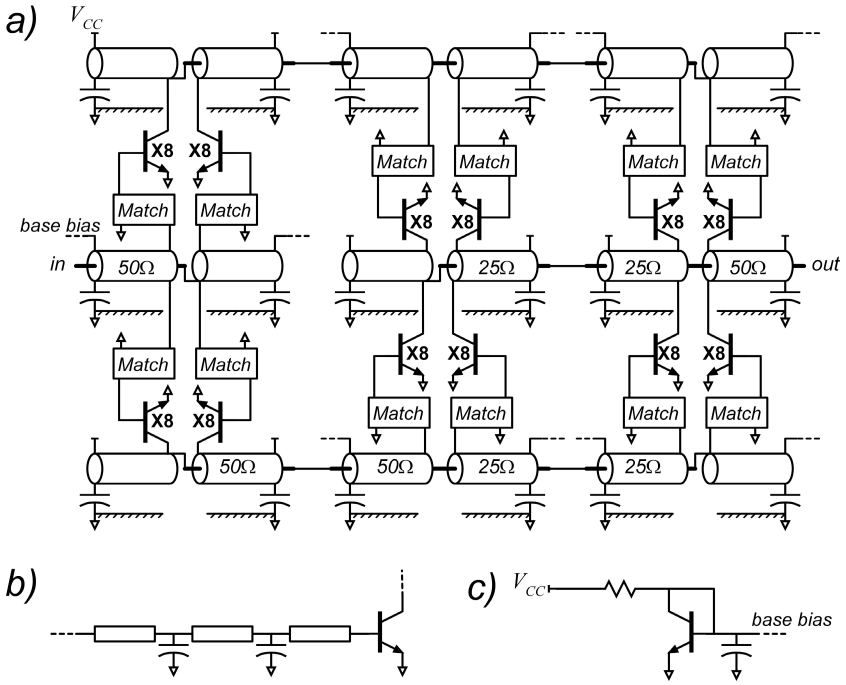

Figure 5.6: Circuit Diagram of 2-stage power amplifier

5.6 Ultra-wideband Power Amplifier Highside Injection Design

A second wideband power amplifier design was created with the intention of creating a receiver with highside local oscillator injection. The amplifier was designed to have high output power from 94 GHz to 165 GHz. Similar to the previous design, a 4:1 2-way combining sub-$\frac{\lambda}{4}$ baluns was used as a power combiner. The increased design frequency reduced the gain which was available from each transistor stage. To reduce this effect, there were two gain stages placed between the input and output balun. This allows the gain of the amplifier stages to overcome the insertion loss of the baluns. Additionally, the baluns were equipped with transmission line sections between balun ports in order to create space to place the transistor cells closer to the balun. This reduces the series inductance between the amplifier and the balun. A screen-capture of high-side injection

local oscillator driver amplifier design can be seen in figure 5.7. This amplifier does not have a driver stage outside of the baluns, so a Wilkinson splitter was used to provide an in-phase signal to each port. A Wilkinson splitter test-structure with a 50Ω termination on one side was also included to de-embed the amplifier performance from the Wilkinson splitter insertion loss.

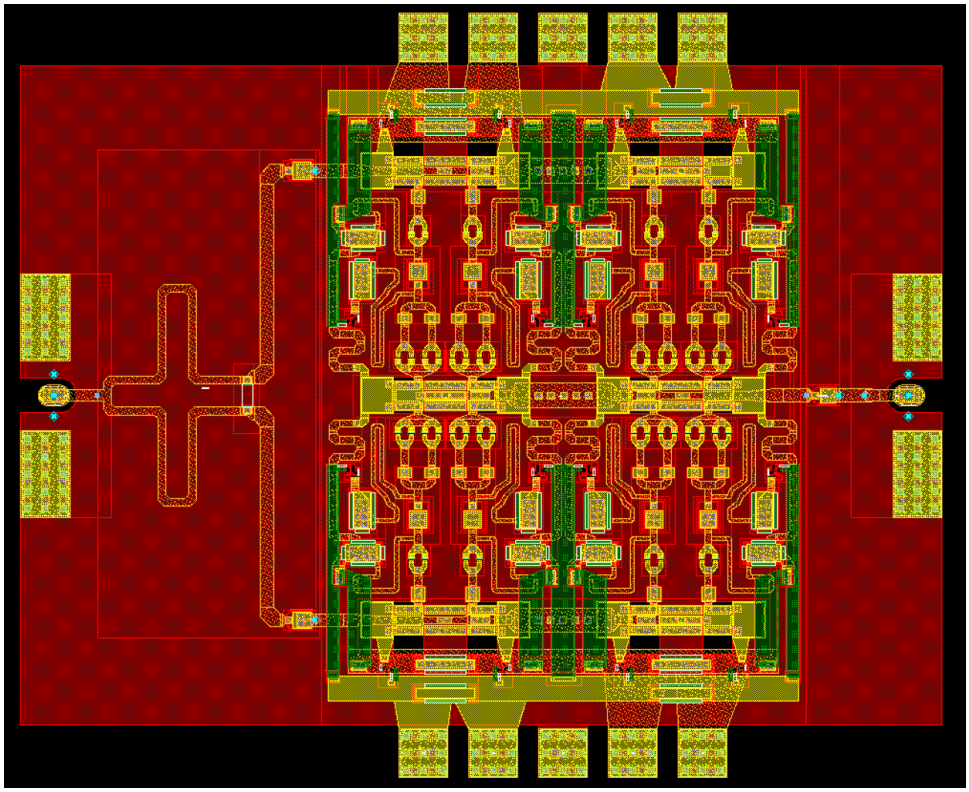

Figure 5.7: Screencapture of high-side injection power amplifier design

The amplifier simulations show a small signal gain of 10 dB at 120 GHz, and a 3-dB bandwidth extending from 100 GHz to 160 GHz. Similar to the low-side injection design, S_{22} is expected to be below -10 dB from 100 GHz to above 170 GHz due to the low device base-collector capacitance and the power combining method. S_{11} is projected to be below -5 dB from 80 GHz to 170 GHz. The 3-dB compression point is simulated to be above 22 dBm from 105 GHz to 165 GHz, with an output power of 23.6 dBm at 125 GHz at

3-dB gain compression. The power added efficiency is simulating to be above 8% from 105 GHz to 165 GHz, including a maximum power added efficiency of 14% at 125 GHz. To our knowledge, these numbers would represent world record PAE at all frequencies from 130 GHz to 165 GHz, and world-class large-signal bandwidth. This IC is expected to be returned in May or June, at which time it will be measured.

5.7 High-Linearity Amplifier Measurements

The ICs were tested for small-signal gain using an Agilent 8510XF Vector Network Analyzer (VNA). Probe tip LRRM calibration was performed with WinCal XE on a Cascade Microtech calibration substrate. The reference plane was set at the probe pads of the IC. The small signal gain of the amplifier is shown in figure 5.8a. For these measurements, $I_c = 100mA$ and $V_{cc} = 2.0V$. The de-embedded small-signal gain of the pseudo-differential amplifier is 6.2 dB at 100 GHz.

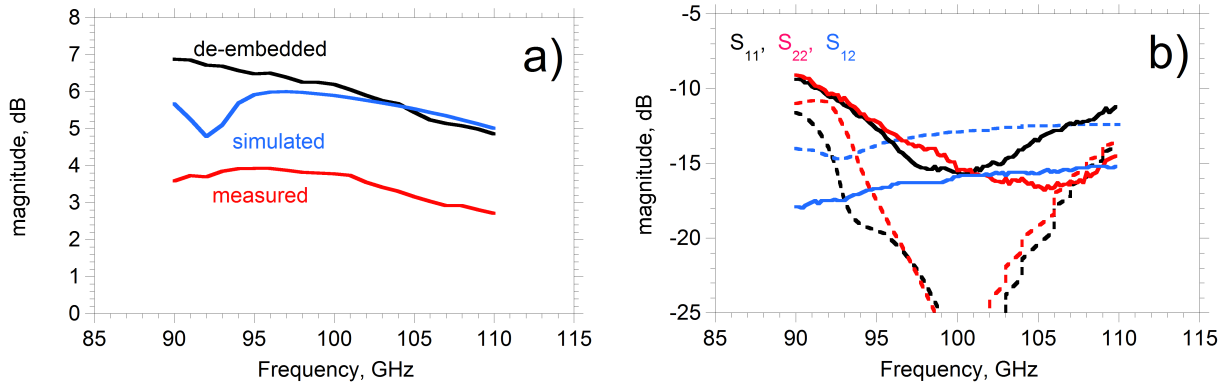

Figure 5.8: a) measured, simulated, and de-embedded S_{21} for 100 GHz high-linearity amplifier and b) measured and simulated S_{11}, S_{22}, and S_{12} for full amplifier breakout (not de-embedded)

Noise measurements were performed using a Micronetics Inc. 90-95 GHz noise source and an Agilent N8972A Noise Figure Analyzer (NFA). The NFA was limited to a max-

imum input frequency of 1.5 GHz, therefore a QuinStar Technology W-band balanced mixer was used to downconvert the output to a compatible frequency. The LO was supplied by a QuinStar 94 GHz Gunn oscillator. Because of the narrow bandwidth of the noise source and the NFA, noise measurements were limited to a frequency range from 94 GHz to 95 GHz. The measured and simulated noise figures of the amplifier can be seen in figure 5.9.

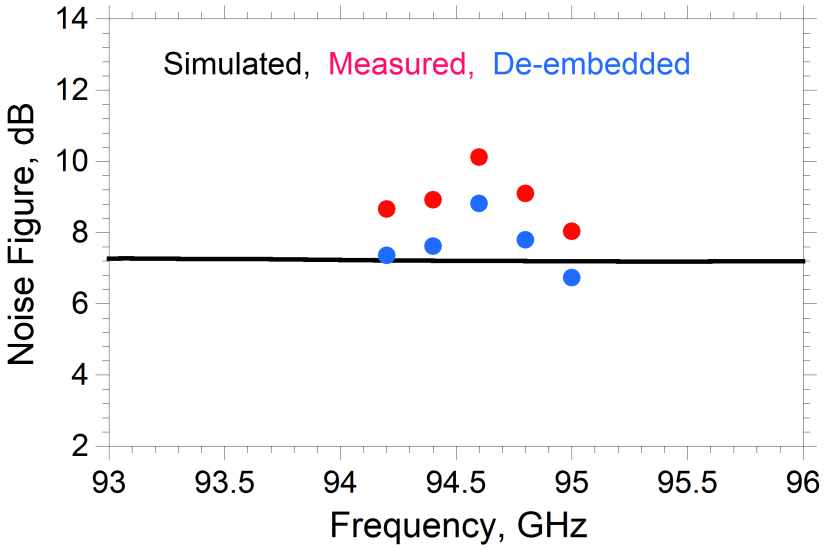

Figure 5.9: Simulated, measured, and de-embedded noise figure for high-linearity amplifier

The two-tone IP3 measurement system was constructed using WR-10 waveguide components. A block diagram of the setup can be seen in figure 5.10. One fundamental tone was produced by an Agilent N5242A PNA-X network analyzer with an OML W-band extender head. A second fundamental tone spaced 100 MHz above the first was produced by an Agilent N5247A PNA-X network analyzer with a W-band active frequency tripler. The input power of each fundamental tone was controlled using a variable attenuator. The two fundamental tones were combined with a Quinstar WR-10 matched hybrid tee (magic tee) and coupled to the amplifier with a pair of Picoprobe W-band GSG probes.

The output was connected to a -20 dB directional coupler with the through-port con-

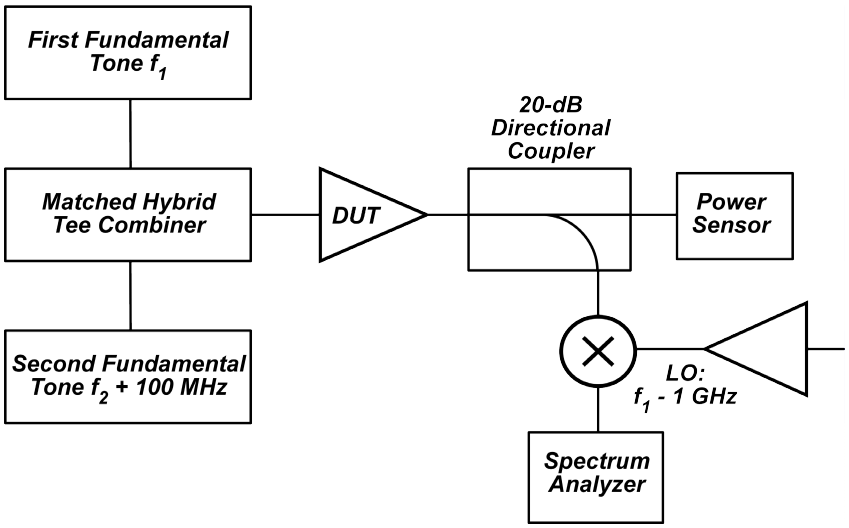

Figure 5.10: Two-tone W-band IP3 measurement configuration

nected to a power sensor to monitor power. The 20dB coupled port was connected to the RF port of a Quinstar W-band balanced mixer. The LO supplied to the downconversion mixer was kept 1 GHz below the first fundamental tone. The IF output was detected using a Rode and Schwartz Spectrum Analyzer. The power of the fundamental tones displayed at 1 GHz and 1.1 GHz and the power of the third-order IM products displayed at 0.9 GHz and 1.2 GHz were recorded for various input powers and used to determine the IP3. The insertion loss of each waveguide component and directional coupler, the conversion loss of the mixer, and the IP3 of the mixer were all measured and used to extract the OIP3 of the amplifier. These measurements were performed at 4 different frequencies at a bias condition of $V_{cc} = 2V$ and $I_c = 100mA$.

The measured S_{21} of the amplifier with baluns and the measured insertion loss of the baluns were then used to de-embed the IIP3 and OIP3 of the core differential amplifier. The IIP3 and OIP3 of the breakout and the de-embedded amplifier can be seen in figure 5.11. The simulation results shown in these plots correspond to those of the core differential amplifier.

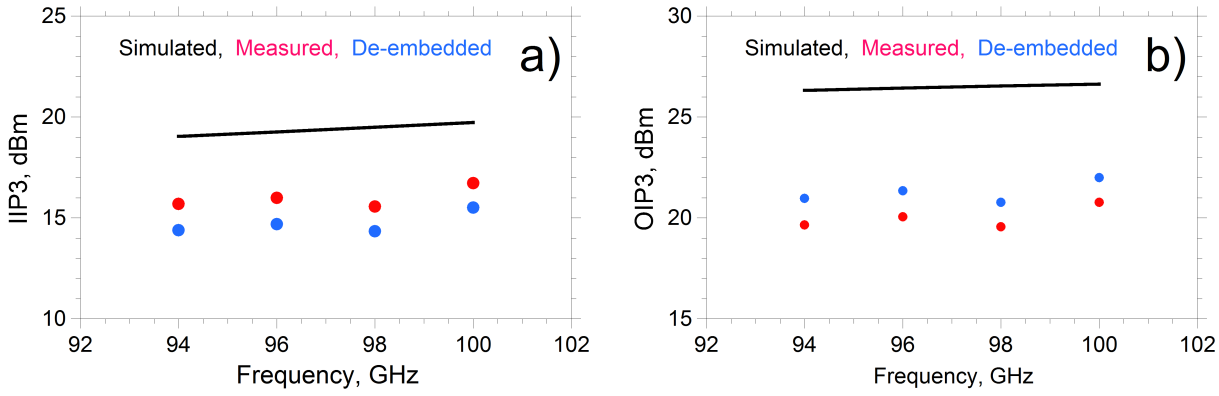

Figure 5.11: measured, simulated, and de-embedded a) IIP3 and b) OIP3

5.8 Ultra-wideband Power Amplifier Measurements

The IC was tested for small signal gain at Teledyne Scientific Company. An Agilent 8510XF Vector Network Analyzer (VNA) was used to perform S-parameter measurements. Probe tip LRRM calibration was performed with WinCal XE on a cascade calibration substrate. The reference plane was set at the probe pads of the IC.

The small signal gain of the 2-stage amplifier is shown in figure 5.12a. For these measurements, $I_c = 360mA$ and $V_c = 2.1V$. As shown in figure 5.12, the small signal gain at 60 GHz is 15 dB. The 3 dB bandwidth extends from 26 GHz to 114 GHz with a local maximum of 18 dB at 30 GHz. Figure 5.12b shows that the measured S_{11} and S_{22} data has been shifted upwards in frequency relative to the corresponding simulated data. We believe that this difference is caused by an error in modeling the IC interconnects, and not an issue with the device model because other IC's designed in the same technology have demonstrated large-scale integration with very little difference between simulated and measured performance.

W-band output power measurements were taken using an Agilent N5242A PNA-X network analyzer with WR-10 Oleson Microwave Laboratory frequency extending heads,

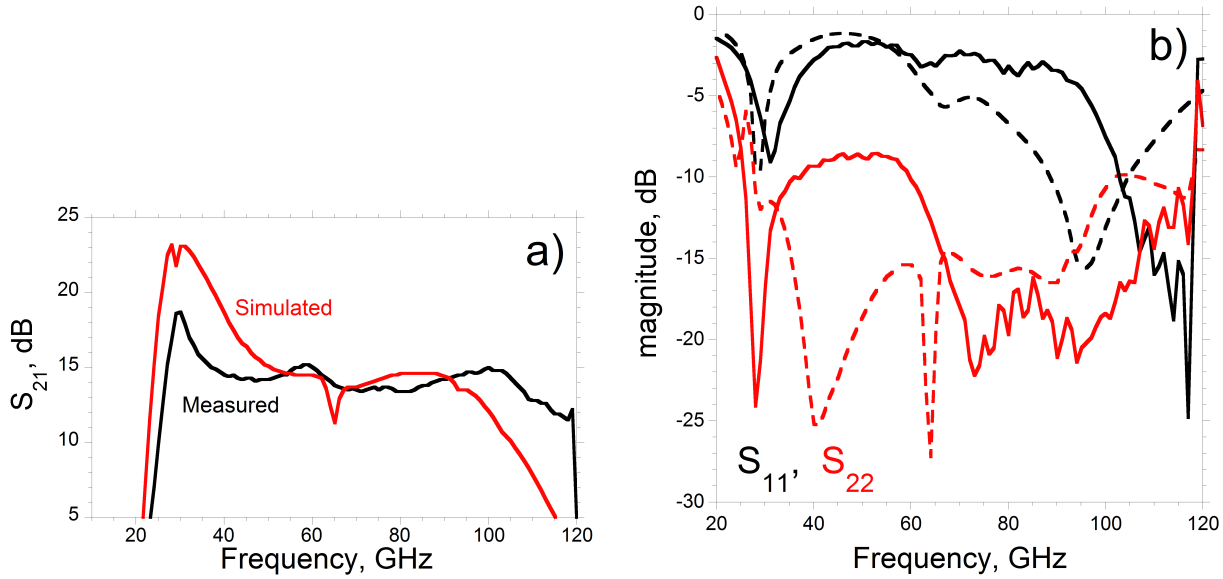

Figure 5.12: a) Measured (black) and simulated (red) S_{21} b) Measured (solid) and simulated (dashed) S_{11} (black) and S_{22} (red). Measured and simulated S_{12} is lower than -30 dB at all frequencies.

followed by a Spacek Labs W-band power amplifier and WR-10 waveguide probes to sweep power at 80 GHz, 90 GHz, and 100 GHz. The resulting output power was measured through an attenuator with an HP power sensor and an agilent E4418B Power Meter. The detected power was calibrated for the insertion loss of the probes, the source chain, and the attenuator. The actual output power vs. nominal output power of the network analyzer and extending heads was also measured and calibrated into the recorded data for each measured frequency.

V-band measurements were taken using an N5247A PNA-X network analyzer to generate an input signal, followed by a passive diode frequency tripler and a Spacek Labs V-band amplifier and a variable attenuator to modulate input drive power. The input drive power was tracked using a 20 dB directional coupler and an HP V-band power sensor and an Agilent E4418B Power Meter. Cascade DC-67 GHz probes were used, and a V-band power sensor and an HP 437B power meter. The attenuation and coupling

factor of the directional coupler and input/output cables were measured and calibrated into the measurement data.

For P_{in} vs P_{out} measurements, the amplifier was biased with $V = 2.24V$ and $I_c = 404mA$. Measurements were made in W-band at 80 GHz, 90 GHz, and 100 GHz. V-band measurements were performed at 50 GHz, 55 GHz, 60 GHz, and 65 GHz. Figure 5.13 shows simulated and measured peak PAE and gain vs. output power at 55 GHz, 90 GHz, and 100 GHz. Figure 5.14 shows peak PAE and output power at the 3-dB compression point vs. frequency.

Note that there is a difference in peak power added efficiency between the simulated and measured data at 80 GHz. We believe this difference is due to an error in modeling the IC interconnects and not an error in the device model itself. We have other ICs in the same technology with a larger number of transistors which show a high level of agreement between the simulated and measured data. Table 5.1 shows a comparison between the work presented here and other high-performance w-band power amplifiers. The low breakdown voltage of the 130nm process limits the maximum output power and power added efficiency, however, to our knowledge, this amplifier demonstrates a wider bandwidth than any previously recorded power amplifier with comprable output power and efficiency.

Ref.	Technology	Freq. (GHz)	BW_{3dB} (GHz)	Max. S_{21}(dB)	P_{out} (dBm)	Peak PAE (%)	V_{dd} or V_{cc} (V)
[31]	$0.25\mu m$ InP HBT	86	23	9.4	20.37	30.4	2.5
[35]	$0.14\mu m$ GaN HEMT	90	35	21	24.5	13.2	12
[36]	$65nm$ CMOS	94	33	18	12	4.5	1.2
[37]	$0.15\mu m$ GaN T-Gate HEMT	91	$\sim$7	16	31.2	20	15
This work	130 nm InP HBT	90	**90**	15	21.95	14.7	2

Table 5.1: Comparison table for high-performance W-band power amplifiers

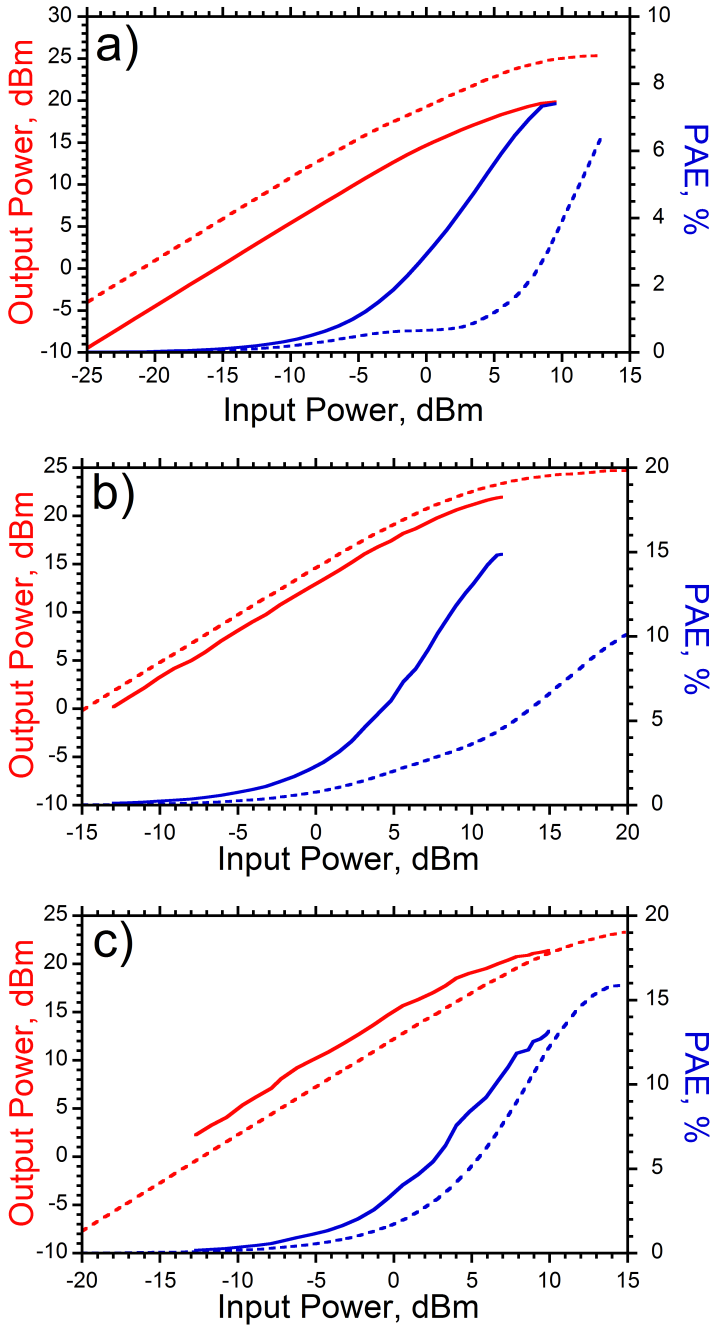

Figure 5.13: Measured (solid) and simulated (dashed) output power (red) and PAE (blue) vs. input power at a) 55 GHz, b) 90 GHz and c) 100 GHz

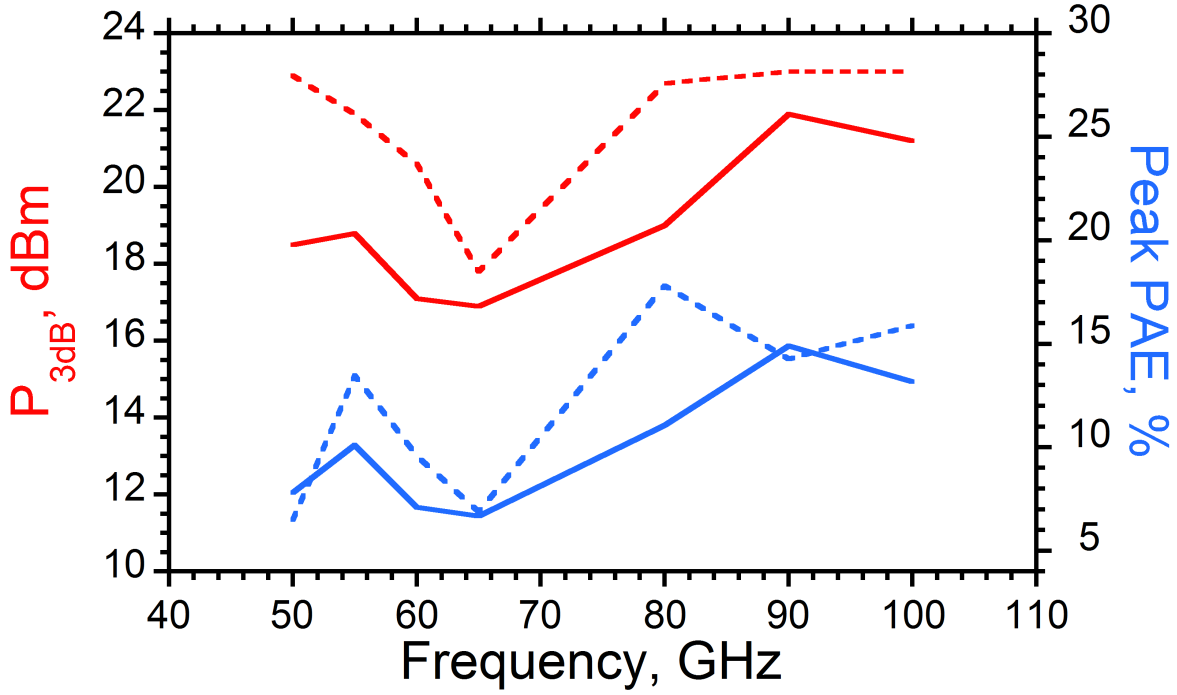

Figure 5.14: measured and simulated 3-dB compression point and Peak PAE vs. frequency

Chapter 6

Future Work and Conclusions

6.1 Future Work

The work presented in this dissertation, particularly the designs for the MIMO spatially multiplexed phased array receiver, will require more work in the future. The expected delivery date for the fabricated beamforming matrix IC breakout has been pushed back from May 2017 to June 2017, the same month of my graduation. Therefore, it is likely that there will be measurements that must still be completed before proceeding with a fully integrated spatially multiplexed multi-beam phased array receiver IC. The current plan is to implement this block into full 4-channel 60 GHz and 140 GHz multi-beam phased array receivers. In order to implement this block into a full receiver, the gain of each pathway must be experimentally determined, as well as the dynamic range of the IC. We would also like to determine the maximum possible data rate of the IC breakout, however, as mentioned before, the breakout measurement data rate may be limited by the wire-bond inductances used to connect the IC to the test board apparatus.

The DARPA ACT project work can also be continued by implementing a high-dynamic range dual conversion receiver with high-side injection of the local oscillator

and with improved dynamic range. The drastically improved linearity of the updated w-band high-linearity amplifier would lead to notably improved numbers for the overall receiver. The high-side injection power amplifier design has reduced gain from the low-side injection design, so more driver amplifiers would be required in order to achieve the same LO drive power. It remains to be seen whether benefits of the reduced in-band spurs from the high-side frequency plan would offset the increased power consumption necessary to drive the LO with a large amount of power.

6.2 Conclusion

Options for expanding wireless network capacities by expanding available bandwidth and increasing spectral efficiency are explored in this dissertation using circuit design techniques. Technologies such as spatially multiplexed multi-beam phased array transceivers can help the commercial wireless industry keep up with the ever increasing demand for wireless data in areas of extreme population density, open city squares, or large event venues. Efficient wide-band mm-wave power amplifiers and high-linearity amplifiers help expand the frequency boundaries for practical wireless applications as well, which incrementally provides more head-room for the FCC to work with when re-distributing frequency space for applications in the future.

The work presented here outlines a proposed architecture for a spatially multiplexed multi-beam phased array receiver, as well as detailed designs for an analog MIMO beamforming matrix IC, which will be measured as soon as the chips are returned from the foundry. The power amplifier ICs presented demonstrate state-of-the-art small signal and large signal bandwidth for high-performance PAs, and the high-linearity amplifiers presented demonstrate some of the first recorded two-tone linearity measurements for W-band amplifiers. Designs for a 100-165 GHz wide-band power amplifier and an im-

proved version of the high-linearity amplifier have also been outlined, and we expect measurements in the near future.

References

[1] T. H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*. Cambridge University Press, Cambridge, second ed., 1988.

[2] A. A. Abidi, *Direct-Conversion Radio Transceivers for Digital Communications*, *IEEE Journal of Solid-State Circuits* **30** (1995) 1399–1409.

[3] B. Razavi, *Design Considerations for Direct-Conversion Receivers*, *IEEE Transactions on Circuits and Systems* **44** (1997) 428–435.

[4] U. Madhow, *Fundamentals of Digital Communication*. Cambridge University Press, Cambridge, first ed., 2008.

[5] D. Pozar, *Microwave Engineering*. Cambridge University Press, Cambridge, third ed., 2005.

[6] B. Razavi, *Fundamentals of Microelectronics*. John Wiley and Sons, Inc., Danvers, MA, 2008.

[7] H. Friisi, *A Note on a Simple Transmission Formula*, *Proceedings of the IRE* **34** (1946) 254–256.

[8] R. L. Olsen, *The aRb Relation in the Calculation of Rain Attenuation*, *IEEE Transactions on Antennas and Propagation* **26** (1978) 318–329.

[9] H. J. Orchard, *Inductorless Filter*, *Electronics Letters* **2** (1966) 224–225.

[10] H. J. Orchard, *Loss Sensitivities in Singly and Doubly Terminated Filters*, *IEEE Transactions on Circuits and Systems* **26** (1979) 293–297.

[11] C. E. Shannon, *Communication in the Presence of Noise*, *Proceedings of the IRE* **37** (1949) 10–21.

[12] F. Rusek, *Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays*, *IEEE Signal Processing Magazine* **30** (2013) 40–60.

[13] M. Rodwell, *Submicron lateral scaling of HBTs and other vertical-transport devices: towards THz bandwidths*, *European GaAs Conference* (2000) 1–4.

[14] S. Y. Kim, *A Low-Power BiCMOS 4-Element Phased Array Receiver for 7684 GHz Radars and Communication Systems*, *IEEE Journal of Solid-State Circuits (JSSC)* **47** (2012) 359–367.

[15] S. H. Choi, *600 GHz InP HBT frequency multiplier*, *Electronics Letters* **51** (2015) 1928–1930.

[16] T. Reed, *48.8 mW Multi-cell InP HBT Amplifier with on-wafer power combining at 220 GHz*, *IEEE Compound Semiconductor Integrated CIrcuit Symposium (CSICS)* (2011) 223–226.

[17] M. Urteaga, *130nm InP DHBTs with ft ¿ 0.52 THz and fmax ¿ 1.1 THz*, *Device Research Conference* (2011) 281–282.

[18] C. Sheldon, *Spatial Multiplexing Over a Line-of-Sight Millimeter-Wave MIMO Link: A Two-Channel Hardware Demonstration at 1.2Gbps Over 41m Range*, *European Conference on Wireless Technology (EuWiT)* (2008) 198–201.

[19] C. Sheldon, *Four-Channel Spatial Multiplexing Over a Millimeter-Wave Line-of-Sight Link*, *IEEE Microwave Symposium (MTT-S)* (2009) 389–392.

[20] C. Sheldon, *A 2.4 Gb/s Millimeter-Wave Link Using Adaptive Spatial Multiplexing*, *IEEE Antennas and Propagation Society International Symposium (APSURSI)* (2010) 1–4.

[21] J. Hoentschel, *From the present to the future: Scaling of planar VLSI-CMOS devices towards 3D-FinFETs and beyond 10nm CMOS technologies; manufacturing challenges and future technology concepts*, *Semiconductor Tchnology International Conference (CSTIC)* (2015) 1–4.

[22] J. Raskin, *SOI technology pushes the limits of CMOS for RF applications*, *Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems (SiRF)* (2016) 17–20.

[23] F. O., *Advanced SOI Technologies: Advantages and Disadvantages*, *International Workshop on Junction Technology* (2006) 200–203.

[24] S. Partridge, *Silicon-on-insulator technology*, *IEEE Proceedings E- Computerts and Digital Techniques* **133** (1986) 106–116.

[25] R. Victor, *Antenna Arrays and Automotive Applications*. Springer, New York, New York, 2013.

[26] H. Hashemi, *A 24-GHz SiGe Phased-Array Receiver - LO Phase-Shifting Approach*, *IEEE Transactions on Microwave Theory and Techniques* **53** (2005) 614–626.

[27] S. Y. Kim, *A 4-bit Passive Phase Shifter for Automotive Radar Applications in 0.13 m CMOS*, *IEEE Compound Semiconductor Integrated Circuit Symposium (CSICS)* (2009) 1–4.

[28] S. Y. Kim, *An Improved Wideband All-Pass I/Q Network for Millimeter-Wave Phase Shifters*, *IEEE Transactions on Microwave Theory and Techniques* **60** (2012) 3431–3439.

[29] L. Zhang, *A 0.1-to-3.1 GHz 4-Element MIMO Receiver Array Supporting Analog/RF Arbitrary Spatial Filtering*, *IEEE Solid-State Circuits Conference (ISSCC)* (2014) 410–412.

[30] W. Ciccognani, *Split gate line distributed power amplifier using tapered drain line approach and active broadband input power divider*, *European Microwave Conference* (2009) 1425–1428.

[31] H.-c. Park, *Millimeter-Wave Series Power Combining Using Sub-Quarter-Wavelength Baluns*, *IEEE Journal of Solid-State Circuits (JSSC)* **49** (2014) 2089–2102.

[32] H.-C. Park, *An 81 GHz, 470 mW, 1.1 mm2 InP HBT Power Amplifier with 4:1 Series Power Combining Using Sub-quarter-wavelength Baluns*, *IEEE Microwave Symposium (IMS)* (2014) 1–4.

[33] H.-C. Park, *30% PAE W-band InP Power Amplifiers using Sub-quarterwavelength Baluns for Series-connected Power-combining*, *IEEE Compound Semiconductor Integrated Circuit Symposium (CSICS)* (2013) 1–4.

[34] S.-K. Kim, *A High-Dynamic-Range W-band Frequency-Conversion IC for Microwave Dual-Conversion Receivers*, *IEEE Compound Semiconductor Integrated Circuit Symposium (CSICS)* (2016) 1–4.

[35] A. Margomenos, *70-105 GHz Wideband GaN Power Amplifiers*, *European Microwave Integrated Circuits Conference* (2011) 199–202.

[36] K. Wu, *77-110 GHz 65-nm CMOS Power Amplifier Design*, *IEEE Transactions on Terahertz Science and Technology* **4** (2014) 391–399.

[37] A. Brown, *W-band GaN Power Amplifier MMICs*, *Microwave Symposium Digest* (2011) 1–4.