Millimeter-Wave Receiver and Package Design Close to the Device Activity Limits

Nima Baniasadi

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2023-40
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-40.html

May 1, 2023

Acknowledgement

I want to express my profound gratitude to my research advisor, Prof. Ali M. Nikne- jad. He provided me with excellent academic guidance and supported me personally, and taught me to be patient when facing various problems. I also thank Prof. Mehrdad Sharif Bakhtiar and Prof. Ali Fotowat Ahmady, my undergraduate professors at Sharif University of Technology, from whom I learned a lot.
This journey would not have been possible without the support of my family and friends. I am grateful to my parents and brother; they have always helped me wherever and whenever I needed them the most. I would also like to thank my friends who provided me with a warm and supportive environment during the years of the pandemic and political crisis.

Millimeter-Wave Receiver and Package Design Close to the Device Activity Limits

by

Nima Baniasadi

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Ali M. Niknejad, Chair
Professor Elad Alon
Professor William L. Holzapfel

Fall 2021

Millimeter-Wave Receiver and Package Design Close to the Device Activity Limits

Abstract

Millimeter-Wave Receiver and Package Design Close to the Device Activity Limits

by

Nima Baniasadi

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Ali M. Niknejad, Chair

For several decades, rapid improvements in the semiconductor industry, particularly the scaling of CMOS processes, have enabled high-speed wireless communications. However, the scaling of CMOS processes seems to be paying off less and less. Moreover, as the carrier frequency increases, the limited power of the CMOS chip can be quickly dissipated by passive elements or at the edges of the chip. The next generations of high-speed radios will require co-design and co-optimization of the chip and package to ensure that the highest data rates are achieved.

This work addresses the design of a packaged wideband millimeter-wave radio. The fundamental limitations of the CMOS process for millimeter-wave applications are examined. Noise measure theory is used to design low-noise amplifiers near the device activity limits. New techniques for minimizing the insertion loss of passive matching networks are proposed. The challenges of a package design are investigated, and an optimized transition structure is proposed. Finally, a 140GHz wideband receiver operating at half the transit frequency of the technology is implemented.

To my parents
for their endless love and support

and to my brother
who made me laugh countless times.

Contents

List of Figures

List of Tables

Acknowledgments

I want to express my profound gratitude to my research advisor, Prof. Ali M. Niknejad. He provided me with excellent academic guidance and supported me personally, and taught me to be patient when facing various problems. I also thank Prof. Mehrdad Sharif Bakhtiar and Prof. Ali Fotowat Ahmady, my undergraduate professors at Sharif University of Technology, from whom I learned a lot.

This journey would not have been possible without the support of my family and friends. I am grateful to my parents and brother; they have always helped me wherever and whenever I needed them the most. I would also like to thank my friends who provided me with a warm and supportive environment during the years of the pandemic and political crisis.

Chapter 1

Introduction

1.1 Connectivity

The internet has become an essential part of daily life over the last three decades. The need for a stable Internet connection is so high that it can be considered an additional layer to Maslow's hierarchy of human needs (Fig. 1.1). For example, during the COVID-19 crisis, the internet played a crucial role in keeping people connected despite physical isolation.

Although the speed of the internet has increased dramatically, user demand has also increased exponentially, and it continues to grow. For example, 5G subscriptions will reach 4.4 billion by 2027 (Fig. 1.2). While Internet Service Providers (ISPs) are upgrading their infrastructure to meet users' needs, they need to keep costs low in such a competitive environment where operators demand 99.999% availability (about 5 minutes of downtime per year) [5].

Subsequent generations of mobile communications have smaller cells for higher spectral efficiency and lower path loss in free space. Fig. 1.3 shows the required backhaul capacity, which is in the tens of Gbps.

Operators have the choice to deploy different technologies. For a detailed comparison of fiber and wireless technologies and their tradeoffs, see Table. 1.1 and Fig. 1.4.

Fiber optic cables are often prohibitively expensive to deploy, and they are still prone to breaks and lengthy disruptions. However, they will be essential for core and inner-city aggregation sites with extremely high capacity requirements. Microwave and millimeter bands are suitable for heterogeneous network backhaul because they allow outdoor cell sites and network aggregation of traffic from multiple base stations, which can then be handed off to mobile switching centers and the core network at the end [6]. Note that tower placement

Figure 1.1: Maslow's hierarchy of human needs with an additional new layer [1]

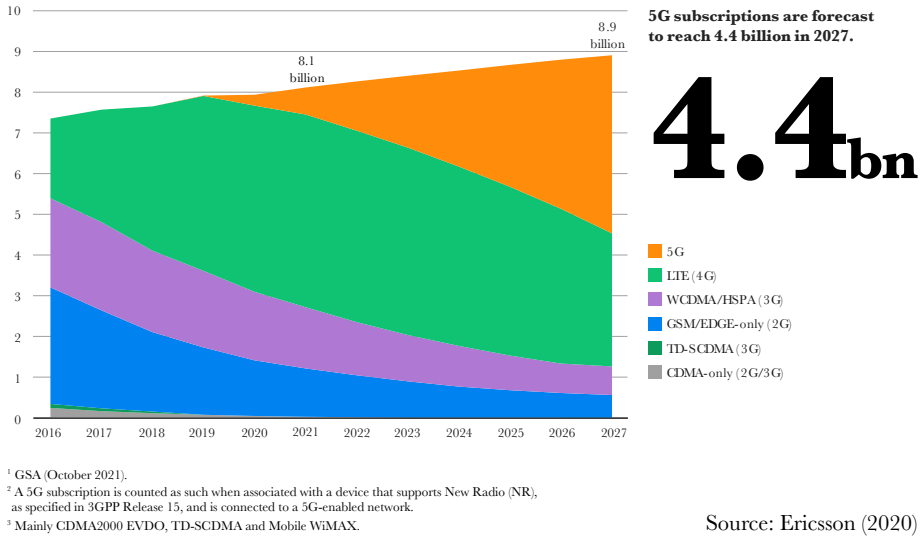


Figure 1.2: Mobile subscriptions by technology (billions)

is not always required in urban areas (antennas can be mounted on rooftops, for example). Therefore, wireless will be used mainly in urban and densely populated areas as the last mile access. It is predicted that between 2021 and 2027, more than 60% of cellular base stations will be connected via microwaves and millimeter waves[7].

Wireless communication in licensed frequency bands increases ISP costs. However, as

Figure 1.3: Backhaul capacity per distributed site

	Wireless	Fiber
Capacity	Up to several Gbps	Unlimited
Regulation	Requires spectrum	Requires right of ways
Deployment Time	Fast deployment time	Increases linearly with distance
Deployment Cost	Increases partially with distance	Increase linearly with distance
Terrain	Requires line-of-sight between two end-points	Costly when trenching in difficult terrain (if accessible)
Climate	Prone to weather conditions	Normally, not affected

Table 1.1: Microwave and fiber consideration [4]

the carrier frequency increases, spectrum costs decrease (Fig. 1.6). On the other hand, the cost of equipment increases as the carrier frequency increases. However, new semiconductor technologies and novel circuit designs reduce these costs. Therefore, using higher carrier frequencies is cost-beneficial to the ISP and subsequently to the end-user.

1.2 Capacity

In this section, the relationship between the link's capacity and the carrier frequency is investigated. Based on Shannon's theorem, the channel capacity C is given by

$$C = \frac{1}{\ln 2} B \ln (1 + SNR) \tag{1.1}$$

Mobile Backhaul Technology Trade-Offs

Wireless vs Fixed vs Satellite

Segment	Microwave (7–40 GHz)	V-Band (60 GHz)	E-Band (70/80 GHz)	Fiber-optic	Copper (Bonded)	Satellite
Future-Proof Available Bandwidth	Medium	High	High	High	Very Low	Low
Deployment Cost	Low	Low	Low	Medium	Medium/High	High
Suitability for Heterogeneous Networks	Outdoor Cell-Site/Access Network	Outdoor Cell-Site/Access Network	Outdoor Cell-Site/Access Network	Outdoor Cell-Site/Access Network	Indoor Access Network	Rural only
Support for Mesh/Ring Topology	Yes	Yes	Yes	Yes where available	Indoors	Yes
Interference Immunity	Medium	High	High	Very High	Very High	Medium
Range (Km)	5~30, ++	1~	~3	<80	<15	Unlimited
Time to Deploy	Weeks	Days	Days	Months	Months	Months
License Required	Yes	Light License/ Unlicensed	Licensed/ Light License	No	No	No

Note: Shading indicates preferred choice for 5G mobile backhaul.

Source: ABI Research

Figure 1.4: Mobile backhaul technology trade-Offs

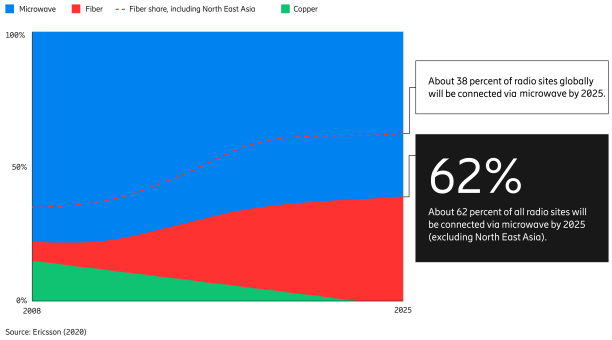


Figure 1.5: Global backhaul media distribution

Assuming a white profile for thermal noise, $SNR = \frac{P_r}{\kappa B}$ and the capacity is

$$C = \frac{1}{\ln 2} B \ln\left(1 + \frac{P_r}{\kappa B}\right) \tag{1.2}$$

where κ is the background noise level. It is not immediately clear whether increasing the total bandwidth contributes to the increase in channel capacity or not, since a higher bandwidth allows for a higher thermal noise. Further investigation of this relationship,

$$\frac{\partial C}{\partial B} = \ln\left(1 + \frac{P_r}{\kappa B}\right) - \frac{1}{1 + \frac{\kappa B}{P_r}} \tag{1.3}$$

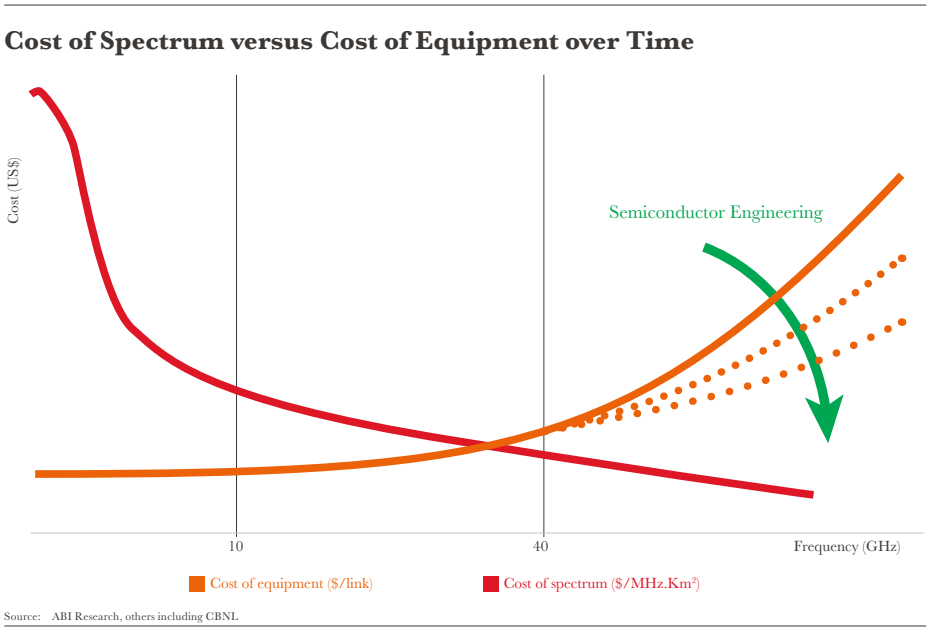


Figure 1.6: Cost of spectrum vs. cost of equipment over time

shows that increasing the absolute bandwidth always increases the capacity, since $\frac{\partial C}{\partial B} > 0$. Since B represents the absolute bandwidth, the same capacity can be obtained for different carrier frequencies (f_c). Defining the fractional bandwidth as

$$B_F = \frac{B}{f_c} \tag{1.4}$$

most radio systems support a limited fractional bandwidth. There are several reasons for this, to name a few:

- Despite the existence of ultra-wideband antennas, most high-efficiency antennas have a relatively limited fractional bandwidth.
- High-frequency circuits tend to use resonators to compensate for the parasitic capacitance of the various elements. The Bode-Fano criterion [8] places an upper limit on the achievable bandwidth when parasitic reactive elements are present.

Therefore, it seems reasonable to use higher carrier frequencies to achieve higher capacity at a given fractional bandwidth (Fig. 1.7). However, as explained in the next section, it should be kept in mind that power generation at higher frequencies is less efficient, and the generated power is attenuated when propagating through the air. Therefore, THz radio systems use phased arrays to generate higher power. Note that phased arrays can increase

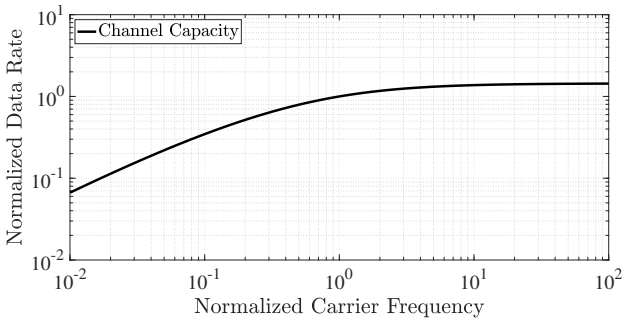


Figure 1.7: Capacity vs. carrier frequency

the directivity of the radiation compared to other power combining techniques, resulting in a higher EIRP[1].

In the non-ionizing frequency range of the electromagnetic spectrum, safety protocols [2] limit the output power of each radiator to avoid electrostimulation of nerve and muscle cells (mainly below 1MHz) or excessive tissue heating. Based on Fig. 1.8, the power density (P_D) is defined as

$$P_D = \frac{P_t G_t}{4\pi d^2} \tag{1.5}$$

should be less than $10\mathrm{W\,m^{-2}}$, where P_t is the transmit power and G_t is the antenna gain, and d is the minimum distance in any direction from any part of the radiating structure to the user's body. The FCC[2] currently regulates the maximum EIRP level, which must be below 55dBm/MHz[9, 10] for a wide range of frequencies to ensure user safety. While service providers should adhere to this limit, power generation in the millimeter-wave band becomes extremely difficult, and most of these systems have limited total output power. Based on

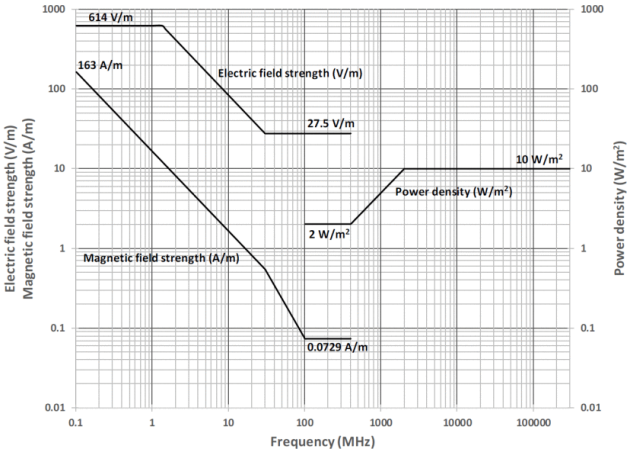


Figure 1.8: Safe radiation levels for persons in unrestricted environments [2].

[1]Equivalent Isotropically Radiated Power
[2]Federal Communications Commission

Friis's formula, the received power is

$$P_r = P_D \frac{\lambda^2 G_r}{4\pi} \tag{1.6}$$

where λ is the wavelength and G_r is the antenna gain of the receiver. The effective area A_e of the receiver is defined as

$$A_e = \frac{\lambda^2 G_r}{4\pi} \tag{1.7}$$

With a continuous wave approximation of $\lambda \approx \frac{v}{f_c}$ and a maximum output EIRP of $P_{max} = P_t G_t|_{max}$, the channel capacity can be written as

$$C = \frac{1}{\ln 2} B_F f_c \ln\left(1 + \frac{\frac{P_{max}}{4\pi d^2}\frac{\lambda^2 G_r}{4\pi}}{\kappa B_F f_c}\right) \tag{1.8}$$

$$= \frac{1}{\ln 2} B_F f_c \ln\left(1 + \frac{v^2}{(4\pi)^2}\frac{1}{d^2}\frac{P_{max} G_r}{\kappa B_F}\frac{1}{{f_c}^3}\right) \tag{1.9}$$

Assuming a user device with a single antenna and a fixed fractional bandwidth, the optimal carrier frequency for the maximum channel capacity can be found by solving the following equation.

$$\frac{\partial C}{\partial f_c} = 0 \tag{1.10}$$

With a change in the variables $\mathfrak{X} = 1 + \frac{\mathfrak{S}}{{f_c}^3}$ and $\mathfrak{S} = \frac{v^2}{(4\pi)^2}\frac{1}{d^2}\frac{P_{max} G_r}{\kappa B_F}$

$$C = \frac{1}{\ln 2} B_F \sqrt[3]{\frac{\mathfrak{S}}{\mathfrak{X} - 1}} \ln(\mathfrak{X}) \tag{1.11}$$

and

$$\frac{\partial C}{\partial \mathfrak{X}} = \frac{-1}{\ln 2} B_F \left(\frac{\mathfrak{S}}{3\sqrt[3]{\left(\frac{\mathfrak{S}}{\mathfrak{X}-1}\right)^2}\left(\mathfrak{X}-1\right)^2}\right)\left(\ln(\mathfrak{X}) - 3 + \frac{3}{\mathfrak{X}}\right) \tag{1.12}$$

The maximum capacity can be reached when

$$\mathfrak{X} = \mathrm{e}^{\mathrm{W}_0\left(\frac{-3}{\mathrm{e}^3}\right)+3} \approx 16.8 \tag{1.13}$$

where e is the Euler's number and W(.) is the Lambert function. The important observation here is that for a maximum allowable transmitter EIRP and a fixed fractional bandwidth, the carrier frequency should be increased so that the total SNR at the end of the receive chain is approximately $\approx$ 12dB, suggesting that for a high-speed over-the-air communication (with a target bit error rate of 10^{-3}), low-order digital modulations such as QPSk or 16-QAM should be used [3]. Higher-order modulations increase the spectral efficiency, but the absolute

[3]QPSK also has the advantage of allowing power amplifiers to operate at their saturated power.

bandwidth must be reduced to achieve the same bit error rate, which ultimately lowers the data rate. The maximum channel capacity of

$$C_{max} \approx \sqrt[3]{\frac{4.2v^2}{(4\pi)^2}\frac{1}{d^2}\frac{P_{max}G_r}{\kappa}{B_F}^2} \tag{1.14}$$

is obtained when the carrier frequency is chosen as

$$f_{c,opt} \approx \sqrt[3]{\frac{1}{15.8}\frac{v^2}{(4\pi)^2}\frac{1}{d^2}\frac{P_{max}G_r}{\kappa B_F}} \tag{1.15}$$

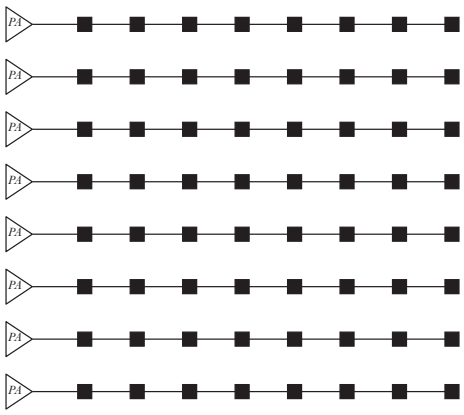


Figure 1.9: An 8×8 2-D transmitter phased array with 1-D steering capability.

Now, let us consider the impact of phased arrays. Fig. 1.9 represents an example of $N \times N$ 2-D phased array that is steerable in one dimension only. In this figure, the black squares represent patch antennas that are series-fed. Assuming that each PA has a maximum output power of P_e per element, the total radiated power is

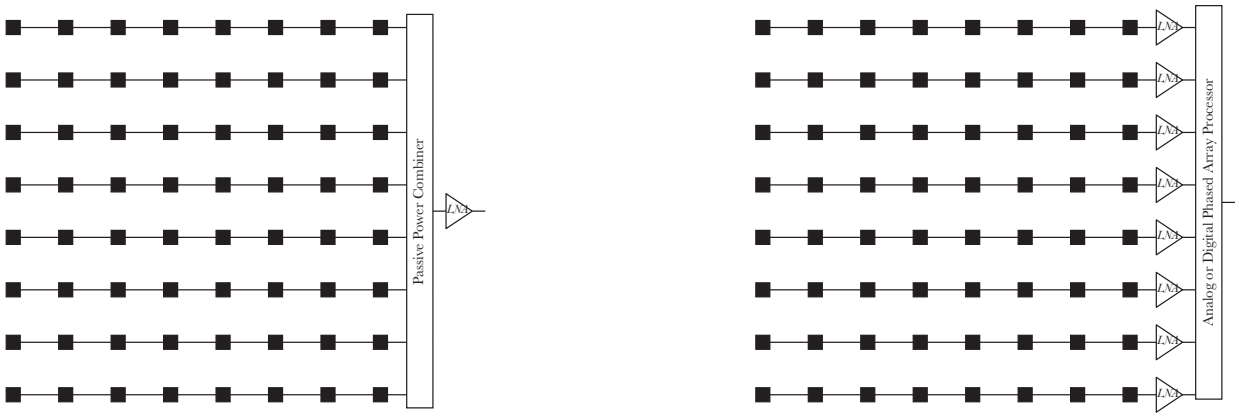
$$P_{max} = N \times P_e \tag{1.16}$$

In practice, the routing loss and the antenna's efficiency should also be considered. As for the directivity of the array, an $N \times N$ array of antennas with $\frac{\lambda}{2}$ spacing provides a directivity of N^2. Moreover, patch antennas provide an additional advantage since these antennas radiate from the front side and ideally have no backside radiation. Hence,

$$G_t = 2N^2 \tag{1.17}$$

which sets the maximum EIPR as

$$P_{max} = 2N^3 P_e \tag{1.18}$$



(a) Passive array processing (b) Active array processing

Figure 1.10: An 8×8 2-D receiver phased array

For example, if each PA has 0dBm of output power, EIRP of 30dBm can be achieved.

The receiver side is a bit more challenging. First, let us consider a fully passive power combining for the phased array as depicted in Fig. 1.10a. Assuming the same patch antenna,

$$G_r = 2N^2 \tag{1.19}$$

Recall that the available noise power of a passive device [4] in thermal equilibrium is equal to kT [8], where k is the Boltzmann constant and T is the absolute temperature. In other words, the use of multiple antennas does not affect the thermal noise power picked up from the ambient blackbody radiation. However, it does increase the directivity of the antennas. It may be difficult to understand why the power level of the radiation noise remains constant despite the combined noise of multiple antennas. The reason is that a passive loss-less power combiner with more than two matched ports does not exist [8]. Therefore, the passive combiner will either partially dissipate or reflect the power to the antennas. Another view is that the thermal noise of the individual elements is generally considered uncorrelated. In contrast, the radiation noise picked up by the different antennas is correlated because it has the same origin, namely the environment. Therefore, different sources can add constructively or destructively after the combiner. The important observation is that the SNR increases by a factor G_r when the antenna array is used. After the LNA, the input-referred noise of the LNA ($\mathfrak{N}_{LNA}$) is added directly to the output. Let us now consider the active array from Fig. 1.10b. In this case, the N uncorrelated noise powers of the LNAs are visible at the output. However, since the signal adds correlatedly in the voltage domain, it is amplified by N^2. The ambient thermal noise at the input of each LNA is preserved as kT because the

[4]Including passive antennas and passive power combiners.

passive structures are in thermal equilibrium. However, since they are partially correlated, the correct method to determine the noise level at the output due to the ambient thermal noise (Fig. 1.11) is to use

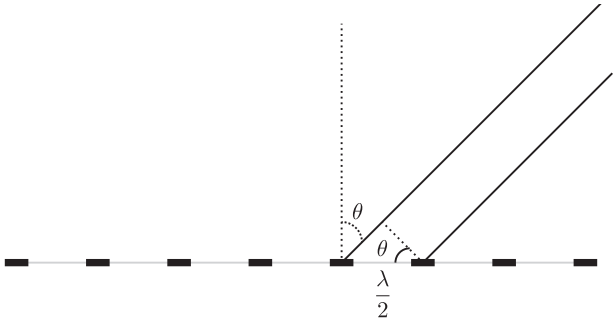


Figure 1.11: 1-D Phased Array

$$V_{\Re,out} = \sum_{m=1}^{N} \left[\iint v_{\Re}(\theta', \phi') \mathrm{e}^{jm\pi \sin(\theta')} g_m(\theta', \phi') \sin(\theta') d\theta' d\phi' \mathrm{e}^{-jm\pi \sin(\theta)} \right] \tag{1.20}$$

where $v_{\Re}(\theta', \phi')$ is the thermal noise source at different spherical locations, $g_m(\theta', \phi')$ is the effective gain of each group of passive antennas for each noise source in spherical coordinates, $\frac{\lambda}{2} \sin(\theta') \frac{2\pi}{\lambda} = \pi \sin(\theta')$ is the phase delay of each noise source to each set of antennas, and $-\pi \sin(\theta)$ is the correction phase that the phased array processor must apply to steer its beam toward the angle of incidence θ. If we assume that each group of antennas is isolated from others (which is not necessarily true [11]), the gain of all groups is approximately equal to $g(\theta', \phi')$ and

$$\overline{V_{\Re,out}{}^2} = \iint \overline{\left(v_{\Re}(\theta', \phi') g(\theta', \phi')\right)^2} \left[\sum_{m=1}^{N} \mathrm{e}^{jm\pi(\sin(\theta') - \sin(\theta))} \right]^2 \sin(\theta') d\theta' d\phi' \tag{1.21}$$

Note that if each group of antennas has a radiation resistance of R,

$$\iint \overline{\left(v_{\Re}(\theta', \phi') g(\theta', \phi')\right)^2} \sin(\theta') d\theta' d\phi' = 4\mathrm{kT}R \tag{1.22}$$

since each group of antennas is in thermal equilibrium. If we assume that ambient thermal noise power coming from different directions is equal ($\overline{v_{\Re}(\theta', \phi')^2} = \overline{v_{\Re}{}^2}$) and each group of series-fed antennas has a uniform radiation distribution in the θ' axis ($g(\theta', \phi') = g(\phi')$), Eq. 1.21 can be simplified to

$$\overline{V_{\Re,out}{}^2} = \int \overline{\left(v_{\Re} g(\phi')\right)^2} \int \left[\sum_{m=1}^{N} \mathrm{e}^{jm\pi(\sin(\theta') - \sin(\theta))} \right]^2 \sin(\theta') d\theta' d\phi' \tag{1.23}$$

Note that the term in the second integral is nothing but the total radiated power (in all directions) of N number of radiators normalized to the total radiated power of a single antenna, equal to N. It follows,

$$\overline{V_{\Re,out}}^2 = 4\mathrm{kT}R \times N \tag{1.24}$$

In other words, there is no difference between passive and active combiners when it comes to the SNR of the received signal due to ambient thermal noise. The only difference is that when multiple LNAs are used in an active combiner, the noise of the LNAs is averaged. For the rest of this section, for simplicity, we consider active combiners with LNA amplification that provides the same signal power as the passive combiner ($G_{P,LNA} = \frac{1}{N}$). Therefore, the noise level of the Shannon's capacity equation is

$$\kappa = \mathrm{kT}\frac{1}{N}\left(N + N\left(\mathcal{F}_{LNA} - 1\right)\right) \tag{1.25}$$

$$= \mathrm{kT}\mathcal{F}_{LNA} \tag{1.26}$$

where $\mathcal{F}_{LNA}$ is the linear noise figure of the LNA. Now, maximum channel capacity and optimum carrier frequency can be calculated as

$$C_{max} \approx \sqrt[3]{\frac{4.2v^2}{(4\pi)^2}\frac{1}{d^2}\frac{4N^5L^2P_e}{\mathrm{kT}\mathcal{F}_{LNA}}{B_F}^2} \tag{1.27}$$

$$f_{c,opt} \approx \sqrt[3]{\frac{1}{15.8}\frac{v^2}{(4\pi)^2}\frac{1}{d^2}\frac{4N^5L^2P_e}{\mathrm{kT}\mathcal{F}_{LNA}B_F}} \tag{1.28}$$

where L models the routing loss and radiation efficiency of the antennas. As a numerical

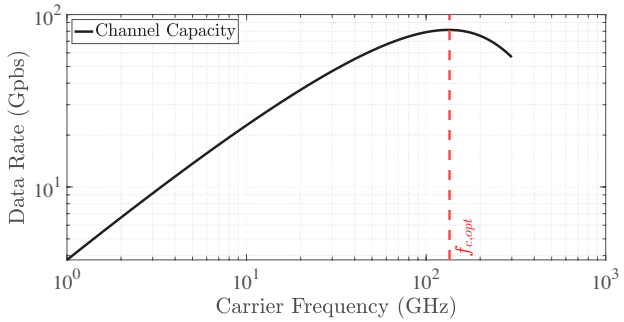


Figure 1.12: Channel capacity vs. carrier frequency for a link based on Table. 1.2

example (Table. 1.2), in an 8×8 1-D steerable phased array where each PA generates 3dBm output power followed by -3dB routing loss, 30dBm EIRP is generated at the transmitter side. Assuming a noise figure of 12dB for the LNAs, a carrier frequency of 135GHz is suitable for 15% fractional bandwidth (20GHz absolute bandwidth) at a distance of 10m. Such a

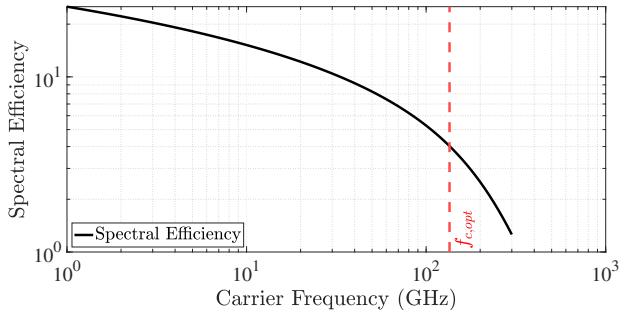


Figure 1.13: Spectral efficiency vs. carrier frequency for a link based on Table. 1.2

Parameter	Value	Description
N	8	Number of antennas in each axis of the TRX phase arrays
P_e	3dBm	Output power of each PA element
B_F	15%	Fractional bandwidth of radio
d	10m	Distance between receiver and transmitter
L	-3dB	Routing loss on PCB
$\mathcal{F}_{LNA}$	12dB	Noise figure of millimeter-wave LNAs
EIRP	30dBm	Equivalent Isotropic Radiated Power
$f_{c,opt}$	135GHz	Optimum carrier frequency for the maximum capacity
C_{max}	81Gbps	Maximum channel capacity

Table 1.2: A numerical example of an optimal over-the-air communication link

system can deliver a data rate of 81Gbps at maximum capacity (Fig. 1.12). It is clear that a simple QPSK modulation is sufficient to achieve the maximum data rate in this system, since SNR has been traded off for a higher data rate (Fig. 1.13). This is easy to understand since

$$C_{max} \approx 4B_F f_{c,opt} \tag{1.29}$$

The strength of phased arrays is easy to see here because $C_{max} \propto \sqrt[3]{N^5}$. However, the physical dimension of the array ultimately limits the number of antenna elements. Let $W = N\frac{\lambda}{2}$ be the width of the antenna array,

$$f_{c,opt} \approx \pi dv \sqrt[2]{2.0\frac{\mathrm{kT}\mathcal{F}_{LNA}B_F}{W^5L^2P_e}v} \tag{1.30}$$

which shows that an optimal carrier frequency should be used for a fixed-size antenna array. The reader should note that the optimal carrier frequency increases as the number of antennas increases as $f_{c,opt} \propto \sqrt[3]{N^5}$, so that the total width of the antenna array is proportional to $W_{opt} \propto \frac{1}{\sqrt[3]{N^2}}$. Although increasing the number of elements without changing the carrier frequency increases the data rate, the channel capacity is suboptimal because a higher carrier frequency increases the data rate. For example, if the antenna array cannot occupy more

than 1cm × 1cm, a carrier frequency of 100GHz with a number of 6 antennas is optimal if the other parameters are taken from Table. 1.2. For a fixed array dimension, the optimal number of elements is

$$N_{opt} \approx \pi d \sqrt{7.9 \frac{1}{W^3} \frac{\mathrm{kT}\mathcal{F}_{LNA} B_F}{L^2 P_e} v} \tag{1.31}$$

So far, increasing the number of elements has increased the channel capacity in the optimal case and made the array dimension smaller. It is easy to see that with this trend, the power density, defined as $\frac{N \times P_e}{W_{opt}} \propto \sqrt[3]{N^5}$, increases with the number of elements used in the array. The high power density makes the packaging of such arrays quite difficult, as they have to cope with higher heat dissipation.

Before concluding this section, let us consider device constraints and their impact on the link capacity. For a CMOS device, the minimum noise figure increases linearly with frequency.

$$\mathcal{F}_{min} = \gamma f_c \tag{1.32}$$

where γ is a technology-dependent proportionality factor. Moreover, the PA survey of

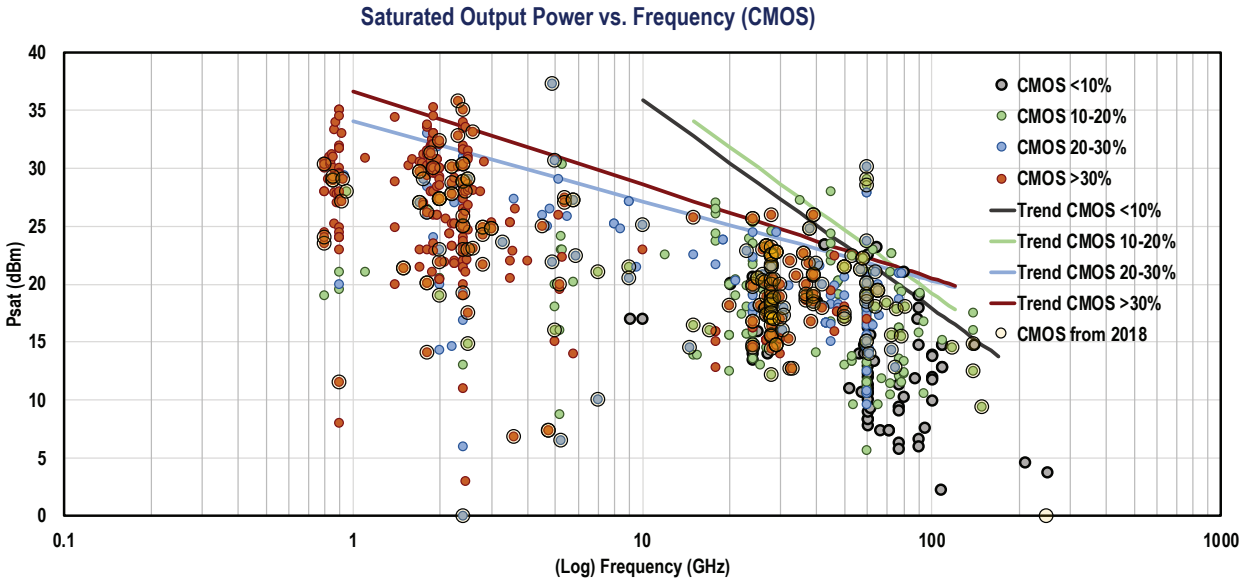


Figure 1.14: The output power of published power amplifiers as a function of carrier frequency[3]

Fig. 1.14 shows that the output power decreases with increasing carrier frequency as

$$P_{e,max} \propto \frac{\mathcal{P}}{f_c} \tag{1.33}$$

where $\mathcal{P}$ is a technology-dependent factor. Now, considering Eq. 1.34

$$W_{opt} \approx \sqrt[5]{2.0\pi^2 \frac{\mathrm{kT}\gamma B_F}{L^2 \mathcal{P}} v^3 \sqrt[5]{d^2}} \tag{1.34}$$

This means that for a given technology and a fixed distance between two transceivers, an optimal array dimension can achieve the maximum data rate of communication.

1.3 Silicon limits

High carrier frequencies require fast transistors. While other compound semiconductors can achieve higher f_t and f_{max}, silicon remains the dominant semiconductor since it has unique capabilities in digital-intense designs. In this section, we will explore some of the limitations of the Bulk CMOS process as shown in Fig. 1.15.

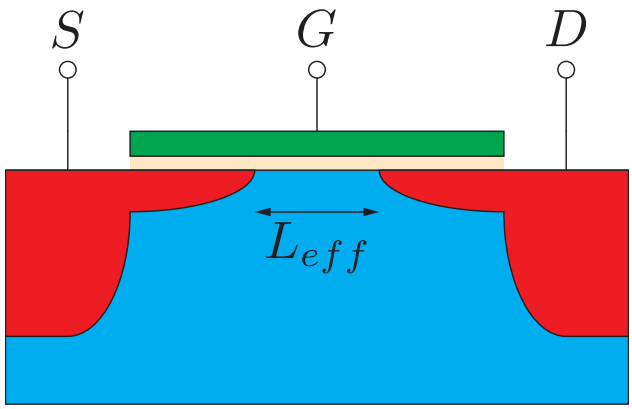


Figure 1.15: A simple model of planar CMOS transistor.

Note that only the semiconductor device is considered here, and the parasitic impact of the back end of the line metallization is not considered. In practice, the performance of deep sub-micron devices is deteriorated by extrinsic parasitic elements.

Transit Frequency f_t

For a CMOS device, the transit frequency f_t is defined as

$$\begin{aligned} f_t =& \frac{1}{2\pi}\frac{g_m}{C_{gs}} \\ =& \frac{1}{2\pi}\frac{\frac{\partial I_{ds}}{\partial V_{gs}}}{\frac{\partial Q_{gs}}{\partial V_{gs}}} \\ =& \frac{1}{2\pi}\frac{\partial I_{ds}}{\partial Q_{gs}} \end{aligned} \tag{1.35}$$

Assuming that the device operates under velocity saturation [5], the maximum drain-source current is reached when all new charges on the source side (∂Q_{gs}) traverse the effective channel length (L_{eff}) at the maximum saturation velocity (v_{sat}). This means

$$\partial I_{ds,max} = \partial Q_{gs}\frac{v_{sat}}{L_{eff}} \tag{1.36}$$

This means that for a CMOS process, there is a maximum limit to the f_t of the device

$$\boxed{f_{t,max} = \frac{1}{2\pi}\frac{v_{sat}}{L_{eff}}} \tag{1.37}$$

To achieve higher current gain and higher transit frequency, either the channel length must be reduced, or the saturation velocity of carriers must be increased. While the latter can be achieved by channel engineering, scaling the channel length remains the main strategy to increase the operating speed of transistors. For example, for a 28nm CMOS node with a saturation velocity of $10^7 \mathrm{cm\,s^{-1}}$, one can expect a maximum f_t of 570GHz. In practice, the effective channel length is about one-third of the drawn channel for the smallest channel length of each node, which can potentially increase the transition speed. On the other hand, the transition frequency is reduced by fringe capacitors and the gate-source and gate-drain overlap capacitors [6], negating any potential improvement from the smaller effective channel length. Once the parasitic capacitance of the back-end metallization is added, the transition frequency drops again.

Analog Efficiency $f_t\frac{g_m}{I_d}$

Another commonly used metric for CMOS transistors is $f_t\frac{g_m}{I_d}$. It gives the analog efficiency of a transistor at a fixed current consumption. Although this metric is not useful for millimeter-wave circuit design, it is still instructive to examine the limits of this metric for a square-law

[5] Diffusion currents are neglected.

[6] For 28nm devices, the fringe capacitors of gate-drain and gate-source are about the same size as the channel capacitance, reducing the transition frequency by a factor of about 3.

device. Analog efficiency can be described as

$$\begin{aligned} f_t \frac{g_m}{I_d} =& \frac{1}{2\pi} \frac{g_m}{C_{gs}} \frac{\frac{2I_{ds}}{V_{od}}}{I_{ds}} \\ =& \frac{1}{\pi} \frac{g_m}{C_{gs}} \frac{1}{V_{od}} \end{aligned} \tag{1.38}$$

where V_{od} is the overdrive voltage. Substitute $g_m = \mu C_{ox} \frac{W}{L} V_{od}$ and $C_{gs} = \frac{2}{3} C_{ox} W L$ into the previous equation, the analog efficiency can be calculated as

$$\begin{aligned} f_t \frac{g_m}{I_d} =& \frac{1}{\pi} \frac{\mu C_{ox} \frac{W}{L} V_{od}}{\frac{2}{3} C_{ox} W L} \frac{1}{V_{od}} \\ =& \frac{1}{\frac{2}{3}\pi} \frac{\mu}{L^2} \end{aligned} \tag{1.39}$$

It should now be clear that for a square-law device with a fixed channel length, the maximum analog efficiency is achieved at a gate-source voltage that provides the highest mobility for the charges in the channel. Fig. 1.16 shows simulation results for NMOS and PMOS devices,

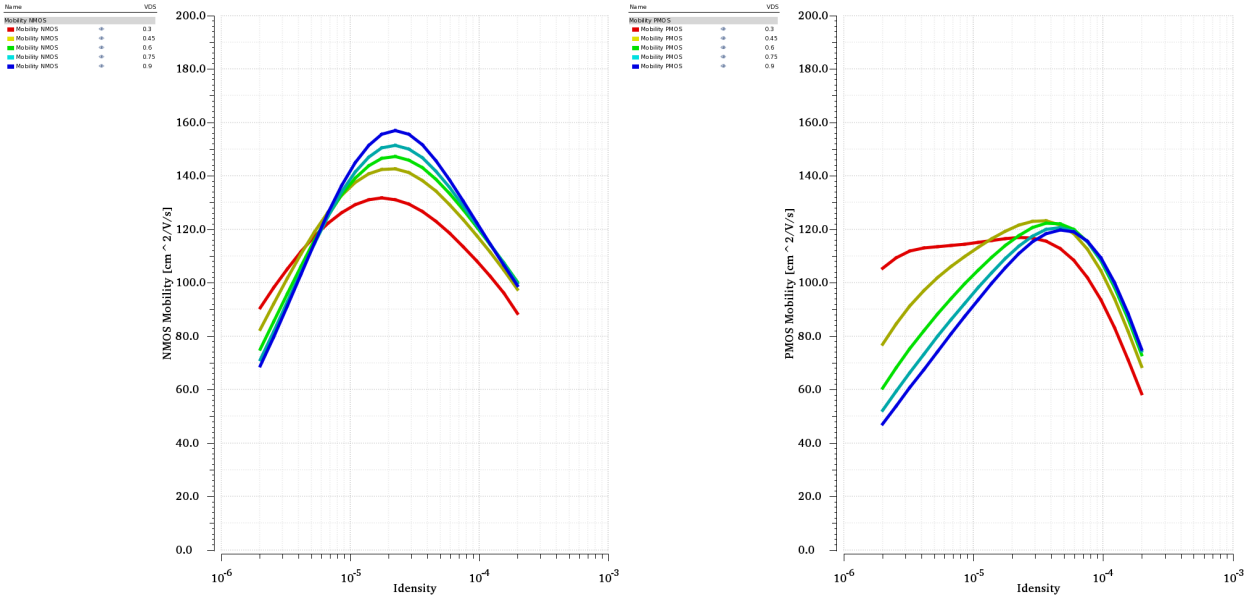


Figure 1.16: Estimating the mobility of the device from simulations for different current densities (A/μm)

where the mobility of the device is given by

$$\begin{aligned} \mu =& \frac{2}{3}\pi f_t \frac{g_m}{I_d} L^2 \\ =& \frac{2}{3}\pi \frac{g_m}{C_{GS} - C_{GD}} \frac{g_m}{I_d} L^2 \end{aligned} \tag{1.40}$$

Note that C_{gs}, the channel capacitance, is replaced by $C_{GS} - C_{GD}$ to remove fringe capacitors. Note that, as expected [12], the NMOS device still performs better than its PMOS counterpart. However, as silicon doping increases, NMOS and PMOS devices become more similar.

Speed-Power Trade-off

Johnson has shown [13] that for bipolar transistors, there is a relationship between the maximum current, the input impedance of the device, and the cutoff frequency of the device. The cutoff frequency is defined by $f_T = \frac{1}{2\pi}\frac{1}{\tau}$, where τ is the average time required for a carrier to traverse the base at the saturated drift velocity. This definition agrees well with the maximum transit frequency defined earlier

$$f_T = f_{t,max} \tag{1.41}$$

Considering a CMOS process with a dielectric breakdown field of E_{Si}, the maximum drain-source voltage can be described as

$$V_{ds,max} = E_{Si} L_{eff} \tag{1.42}$$

Thus, there is a relationship between the maximum drain-source voltage and the device cutoff frequency where

$$V_{ds,max} f_{t,max} = \frac{1}{2\pi} E_{Si} v_{sat} \tag{1.43}$$

A transistor with a transit frequency of 400GHz cannot generate more than 2 volts peak-to-peak drain-source voltage, assuming $E_{Si} \approx 5 \times 10^5 \mathrm{V\,cm^{-1}}$ and $v_{sat} \approx 10^7 \mathrm{cm\,s^{-1}}$.

While the dielectric breakdown field sets a maximum drain-source voltage, any drain-source current can be achieved at the cost of increased input capacitance by connecting multiple devices in parallel. The load current through a charge control device is defined by

$$I_{ds} = \frac{Q_{ch}}{\tau} \tag{1.44}$$

where Q_{ch} is the total mobile charge in the channel, and τ is the average charge transit time. To calculate the maximum current, we should assume the highest drift velocity, which sets $\tau_{min} = \frac{v_{sat}}{L_{eff}}$. Also, we assume a dielectric breakdown field of E_{ox} for the dielectric barrier between the gate and the channel,

$$Q_{max} = C_{gs} E_{ox} t_{ox} \tag{1.45}$$

where t_{ox} is the oxide thickness. It follows that,

$$I_{ds,max} = C_{gs} E_{ox} t_{ox} \frac{v_{sat}}{L_{eff}} \tag{1.46}$$

If we define $X_{f_T} = \frac{1}{2\pi f_T C_{gs}}$ as the reactive input impedance between gate and source, it is clear that

$$I_{ds,max} X_{f_T} = E_{ox} t_{ox} \tag{1.47}$$

In other words, for a fixed impedance at the device cutoff frequency, the maximum drain-source current is fixed by the properties of the gate oxide. For most CMOS process nodes, the drain-source and gate-source breakdown fields are close to each other because the drain of the preceding transistors directly controls the gate nodes of digital circuits. For example, with $E_{ox} = 14\text{MV}\,\text{cm}^{-1}$ for silicon dioxide [14] and an oxide thickness of 1nm,

$$E_{ox} t_{ox} \approx E_{Si} L_{eff} \approx 1.5\text{V} \tag{1.48}$$

For a CMOS process optimized for digital circuits,

$$\begin{aligned} I_{ds,max} =& C_{gs} E_{ox} t_{ox} \frac{v_{sat}}{L_{eff}} \\ \approx& C_{gs} E_{Si} L_{eff} \frac{v_{sat}}{L_{eff}} \\ \approx& C_{gs} E_{Si} v_{sat} \end{aligned} \tag{1.49}$$

Therefore, the relationship between the volt-ampere product (as an approximation for the maximum output power), the input impedance level ($X_f = \frac{1}{2\pi f C_{gs}}$), and the cutoff frequency of the transistor can be found as

$$\begin{aligned} f \times I_{ds,max} \times V_{ds,max} =& f \frac{1}{2\pi} \frac{E_{Si} v_{sat}}{f_T} C_{gs} E_{Si} v_{sat} \\ =& 2\pi f C_{gs} \left(\frac{E_{Si} v_{sat}}{2\pi} \right)^2 \frac{1}{f_T} \end{aligned} \tag{1.50}$$

and thus,

$$f \times I_{ds,max} \times V_{ds,max} \times X_f = \left(\frac{E_{Si} v_{sat}}{2\pi} \right)^2 \frac{1}{f_T} \tag{1.51}$$

Note that

- increasing the cutoff frequency decreases the output power for a fixed technology and a fixed input impedance, and
- for a fixed input impedance, faster process nodes provide lower output power at a fixed operating frequency.

Termination Levels vs. Frequency

The optimal large-signal termination load for maximum power transfer can be studied by finding the proper ratio of $\frac{V_{ds,max}}{I_{ds,max}}$

$$\begin{aligned} R_{opt} = \frac{V_{ds,max}}{I_{ds,max}} =& \frac{\frac{1}{2\pi}\frac{E_{Si}v_{sat}}{f_T}}{C_{gs}E_{Si}v_{sat}} \\ =& \frac{1}{2\pi f_T C_{gs}} \end{aligned} \tag{1.52}$$

which shows that for a fixed input impedance level, output terminations should be smaller for faster process nodes

$$R_{opt} = X_f \frac{f}{f_T} \tag{1.53}$$

Large Signal Power Gain vs. Frequency

In the simple model presented above, gate impedance is considered purely imaginary. It is generally not the case as the operating frequency of the transistor increases since one must consider the non-quasistatic model for the device. First, consider a resistor R_g in series with the gate capacitance. We will discuss the origins of this resistance later. Let us assume a small resistor,

$$I_{gs} = 2\pi f Q_{ch} \tag{1.54}$$

where Q_{ch} is the total mobile charge in the channel. Thus, the peak power dissipation at the series resistor is

$$P_{in} = \frac{1}{2}\left(2\pi f Q_{ch}\right)^2 R_g \tag{1.55}$$

The maximum output current is

$$I_{ds} = \frac{Q_{ch}}{\tau_{min}} \tag{1.56}$$

and therefore the peak output power is

$$P_{out} = \frac{1}{2}\left(\frac{Q_{ch}}{\tau_{min}}\right)^2 R_L \tag{1.57}$$

where R_L is the terminating resistor. Assume that the terminating resistor $R_L = R_{opt}$ is chosen,

$$P_{out} = \left(\frac{Q_{ch}}{\tau_{min}}\right)^2 \frac{1}{4\pi f_T C_{gs}} \tag{1.58}$$

and the power gain can be calculated as

$$\begin{aligned} G =& \frac{P_{out}}{P_{in}} \\ =& \frac{1}{(2\pi f Q)^2 R_g} \left(\frac{Q_{ch}}{\tau_{min}}\right)^2 \frac{1}{2\pi f_T C_{gs}} \\ =& \frac{2\pi f_T}{(2\pi f)^2} \frac{1}{R_g C_{gs}} \end{aligned} \tag{1.59}$$

Considering only the intrinsic device, $R_g C_{gs}$ is the time constant for the redistribution of the channel charge in response to the gate excitation and can be calculated to be about $\frac{1}{5} \sim \frac{1}{8} \times \frac{1}{2\pi f_t}$ for a square-law device [15]. The exact coefficient depends on the exact charge distribution in the channel. Therefore, we assume that α represents this coefficient, which ranges between $5 \sim 8$ for square-law devices and drops to smaller values (≈ 2) for a uniform charge distribution. It follows,

$$G = \alpha \left(\frac{f_T}{f}\right)^2 \tag{1.60}$$

Interestingly, although the output power of faster process nodes tends to decrease for a fixed input impedance, the large-signal power gain improves when faster transistors are used. Defining f_{max} as the frequency at which the power gain drops to 0dB, we find that.

$$f_{max} \propto f_T \tag{1.61}$$

Detailed Model with Extrinsic Parasitics

Fig. 1.17 shows a complex model of a transistor with the first layer of back-end metallization, and the different values are listed in Table. 1.3. The red components are calculated within the BSIM model, while the blue components extract the parasitic elements in the layout. These parasitic elements include:

- R_{tip} is the resistance from the edge of the via to the edge of the OD definition. In HKMG[7] processes, the work function of the metal is used to set the threshold voltage of the device. Therefore, devices with the same layout but different thresholds may have different series resistance and high-frequency response. While other parasitic components scale with the width of the device, the minimum R_{tip}, which corresponds to the shortest distance between the transistor and the poly contact, is fixed by the process capabilities.

[7]HKMG: High-K Metal Gate

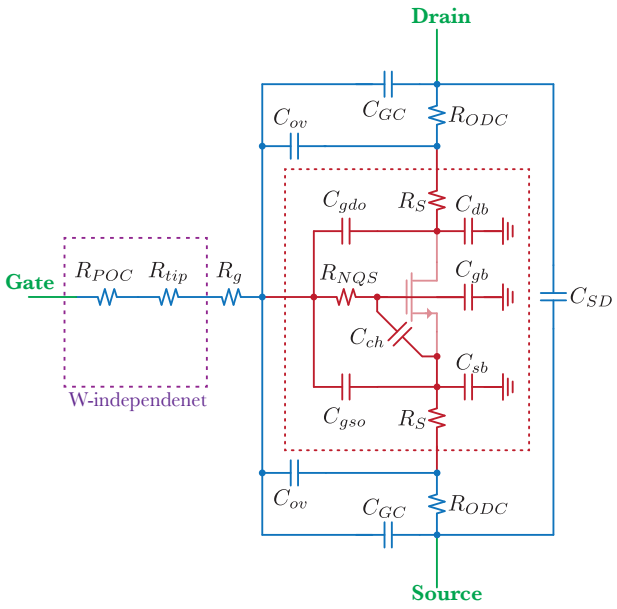


Figure 1.17: Parasitic elements of a single-finger transistor.

- R_{POC} is the poly contact resistance from the first metal layer (M1) to the gate metal (PO). It should be mentioned that some process nodes allow negative enclosure for the contact. It allows devices with a shorter channel length (and consequently faster f_t), but the current crowding at the contact tip increases the contact resistance. Instead of using a single contact, the designer can use multiple contacts in parallel. It seems compelling, but since additional poly contacts are connected in series with the added resistance of the gate extension, the advantage is quickly exhausted. For example, for a single poly contact

$$\begin{aligned} R_{POC} + R_{tip} &= 100 + 117\Omega \\ &= 217\Omega \end{aligned} \tag{1.62}$$

  while two poly contacts

$$\begin{aligned} R_{POC} + R_{tip} &= (100 + 117)||100 + 117\Omega \\ &= 185\Omega \end{aligned} \tag{1.63}$$

  will only provide 15% of improvement. In contrast, using a double-sided contact, as shown in Fig. 1.18, halves the effective resistance $R_{POC} + R_{tip}$ in series with R_g and improves the overall gate resistance by more than %50 [8]. Although the double-sided contact strategy is promising, it requires a complex layout and additional parasitic capacitance. Therefore, the double-sided contact should be used only when necessary.

[8] Since the gate metal is driven from both sides, its effective resistance also decreases

- R_g is the series resistance of the metal gate on the active area (OD). For a planar transistor,

$$R_g = \alpha \rho_{MG} \frac{W}{t_{MG} L} \tag{1.64}$$

  where t_{MG} is the thickness of the metal gate, ρ_{MG} is the resistivity of the metal used for the gate, and $\frac{1}{2}$ assumes a simple T-model [9]. It should be mentioned that when a double-sided contact is used, the effective resistance of the metal gate decreases by a factor of 4.

- C_{ch} is the channel capacitance. It is a nonlinear capacitance that depends strongly on the bias of the gate voltage and the response to large signals.

- C_{gdo} and C_{gso} are the gate-drain and gate-source overlap capacitance modeled in the BSIM model. These are relatively linear, bias-independent capacitors generally symmetric on both the drain and source sides. While these capacitors are negligible for devices with long channels, they become comparable with the highest channel capacitance of the short channel device.

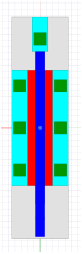
- Similarly, C_{ov} is the stray capacitance between gate to source and drain extracted from the layout extractor. C_{GC} also represents the coupling capacitor between gate and M1.

- C_{sb} and C_{db} represent the source-bulk and drain-bulk junction capacitance, respectively, and C_{gb} represents the gate-bulk capacitance. This capacitor comes from the gate metal extensions running away from the active area. All these capacitors are weakly dependent on the gate bias voltage.

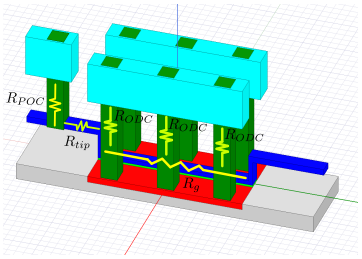
- R_S models the resistance of the shallow drain-source junction extensions. It is one of the most critical limiting factors for the minimum switch resistance in digital circuits [16, 17]. Therefore, it is essential to model this resistance in a passive mixer properly. In analog and high-frequency applications, this physical resistor also generates thermal noise. It acts like a degeneration resistor in series with the device source and ultimately limits the device's transconductance.

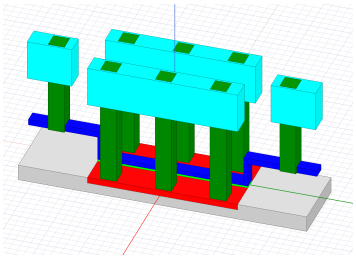
As mentioned earlier, the use of more advanced nodes improves the intrinsic cutoff frequency of the device. However, scaling is detrimental to the effect of back-end metallization. As shown in Fig. 1.19, as the technology scales, not only do the lateral dimensions (e.g., channel length and channel width of the transistors) scale but so do the vertical dimensions (e.g., the thickness of the metals and the thickness of the interlayer dielectrics). If the designer keeps the device's width constant from one technology node to another, driving the same transistor will result in higher power dissipation because the series resistance of the gate has increased. Therefore, despite improving the device's cutoff frequency, the device's power gain may not follow the same improvement. To illustrate this point, let us calculate

[9] Most parasitic extraction programs use a T-model for the gate resistance.



(a) Top view of a single-sided contact





(b) Side view of a single-sided contact (c) Side view of a double-sided contact

Figure 1.18: A simple planar transistor in layout view and its 3D representation

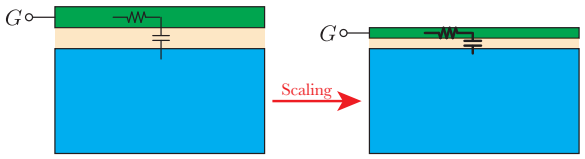


Figure 1.19: Parasitic elements of a transistor.

f_{max} using a simplified model shown in Fig. 1.20. For this unilateral device, the maximum available power gain is

$$G_P = \frac{1}{4} {G_I}^2 \frac{R_{out}}{R_g} \tag{1.65}$$

where $G_I \approx \frac{f_t}{f}$ is the current gain, R_{out} is the device output resistance, and R_g is the total series resistance before channel capacitance. It follows that,

$$\begin{aligned} f_{max} \approx & \frac{f_t}{2} \sqrt{\frac{R_{out}}{R_g}} \\ \approx & \frac{f_t}{2} \sqrt{\frac{g_m R_{out}}{g_m R_g}} \end{aligned} \tag{1.66}$$

If R_g is dominated by the metal gate, assuming a square-law device where $g_m = \mu \frac{\epsilon_{ox}}{t_{ox}} \frac{W}{L} V_{od}$,

$$g_m R_g = \alpha \mu \rho_{MG} \epsilon_{ox} \frac{1}{t_{MG} t_{ox}} \left(\frac{W}{L} \right)^2 V_{od} \tag{1.67}$$

Length	$40nm$	-
Width	$370nm$	-
Nominal Current	$200uA/um$	-
R_{POC}	100Ω	Per "Poly Contact"
R_{tip}	117Ω	Minimum size allowed by DRC
R_g	241Ω	For $370um$ length modeled as T
R_{ODC}	100Ω	Per "OD Contact"
R_S	200Ω	Weakly depends on bias condition ($200 \sim 350\Omega$)
R_{NQS}	400Ω	Strongly depends on bias condition ($\approx \frac{1}{5g_m}$)
C_{SD}	$23.3aF$	For $370um$ length
C_{ov}	$6.6aF$	For $370um$ length
C_{GC}	$9.2aF$	For $370um$ length
C_{gso}	$80aF$	For $370um$ length
C_{gdo}	$80aF$	For $370um$ length
C_{gb}	$30aF$	Weakly depends on bias condition ($15 \sim 35aF$)
C_{sb}	$85aF$	Weakly depends on bias condition ($80 \sim 120aF$)
C_{db}	$85aF$	Weakly depends on bias condition ($80 \sim 120aF$)
C_{ch}	$90aF$	Strongly depends on bias condition ($0 \sim 100aF$)

Table 1.3: Model values.

which is inversely proportional to the scaling trend [10]. Moreover, $g_m R_{out}$ is the intrinsic gain of the device, which decreases proportionally to the scaling factor, assuming a first-order approximation for the channel length modulation. Therefore, despite the improvement of f_t, f_{max} will not follow the same trend and will remain nearly constant unless the mobility of the majority carriers is increased by channel engineering [18] or the conductivity of the metal gate is increased [11]. The significance of this result is that once the extrinsic parasitic elements limit the performance of the device, scaling offers the designer little to no improvement with respect to the device f_{max}.

1.4 Challenges

The previous sections have explained the need for high frequency, high data-rate communication links. While previous works have achieved high-speed links above 100GHz (Fig. 1.21), challenges still exist.

[10] Remember that $g_m R_{NQS}$ would have remained constant if R_g were dominated by the intrinsic gate resistance R_{NQS}.

[11] While reducing ϵ_{ox} seems to be equally effective, it is not desirable because the gate loses its control over the channel

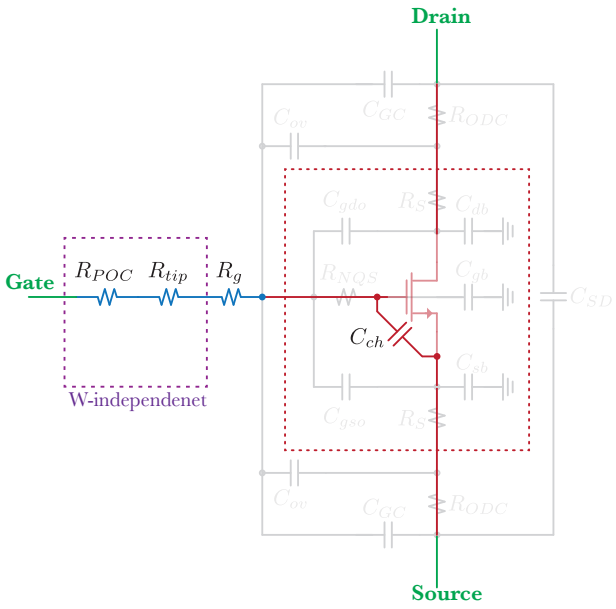
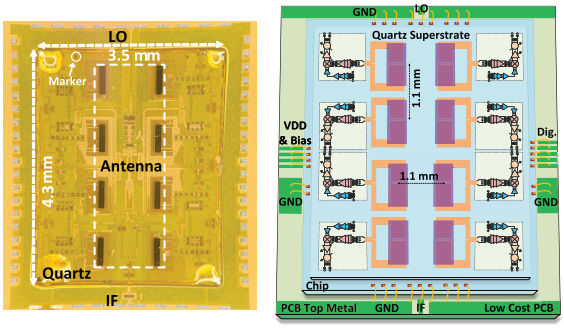


Figure 1.20: Simplified transistor model

The first problem is to increase the link distance from tens of centimeters to several meters. As described in the previous section, using a phased array is effective. However, using a large number of elements increases the system's power consumption. On the other hand, the spacing between elements becomes smaller at high carrier frequencies, which increases the power dissipation density. Therefore, a good packaging approach capable of cooling the various elements of the system should be considered. Also, the cost of the package and the silicon area should be considered together. For example, in Fig. 1.21a, on-chip antennas were used to implement an array of 2×4 elements. Keep in mind that if high directivity antennas were used, a larger portion of the silicon was occupied by passive antennas. Therefore, it makes sense to place the antennas outside the chip. Unfortunately, the signal transition from the chip to the package becomes a challenge at millimeter-wave frequencies, with potentially high insertion loss if not properly designed.

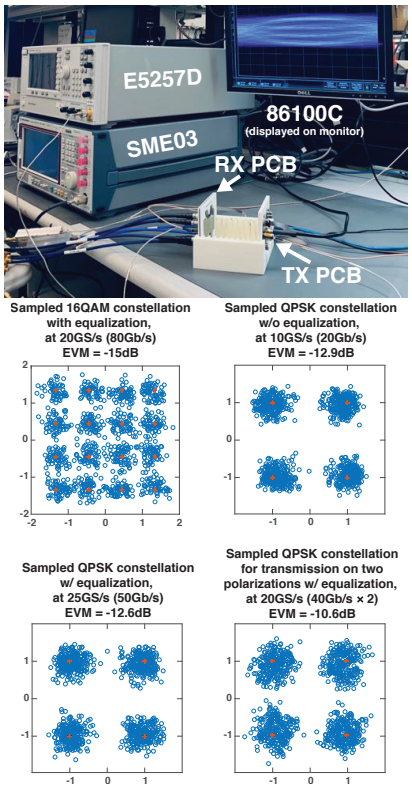
The rest of this article is organized as follows. The next chapter addresses the tradeoff between noise and gain in any amplifier and the noise measure. Chapter 3 looks in detail at the design of a 140GHz wideband receiver. Chapter 4 and Chapter 5 deal with the chip-to-package and inter-package transition of millimeter-wave signals. Chapter 6 concludes this article.



Measured Constellations at Different Data Rates, Scan Angles and Modulations
$P_{inc} = -42$ dBm

Modulation Data rate/EVM	16-QAM 4 Gbps/3.5%	16-QAM 10 Gbps/6.8%	64-QAM 6 Gbps/3.8%	64-QAM 9 Gbps/5.3%
140 GHz 16/64 QAM 0° Scan				
Modulation Data rate/EVM	**16-QAM 4 Gbps/3.7%**	**16-QAM 10 Gbps/6.6%**	**64-QAM 6 Gbps/3.7%**	**64-QAM 9 Gbps/4.7%**
150 GHz 16/64 QAM 0° Scan				
Scan Angle Data rate/EVM	**-30° 3.6 Gbps/4.5%**	**-10° 3.6 Gbps/3.5%**	**10° 3.6 Gbps/3.4%**	**30° 3.6 Gbps/4.3%**
143.5GHz 64 QAM E-plane Scan				

(a) 9Gbps link at 140GHz [19]



(b) 80Gbps link at 115GHz [20]

Figure 1.21: Packaged millimeter-wave radios

Chapter 2

Millimeter-wave LNA Design

2.1 Introduction

The concept of "noise measure" was introduced in [21] by Haus and Adler. It becomes crucial when the operating frequency approachesf_{max} of the active device when the available gain of the device is severely limited. Suppose that a chain of M identical amplifiers with a limited power gain of G is cascaded as shown in Fig. 2.1 to achieve a high power gain. Due to the noise of the amplifiers, the SNR at the output of the chain is degraded by D:

$$D = \frac{\frac{S_{in}}{N_{in}}}{\frac{S_{out}}{N_{out}}} \tag{2.1}$$

where S_{out} and N_{out} are the powers of the output signal and output noise [1] powers and S_{in} and N_{in} are the powers of the input signal and noise, respectively.

$$S_{out} = S_{in} \times G^Q \tag{2.2}$$

$$\begin{aligned} N_{out} &= P_{noise}\left(1 + G + ... + G^{Q-1}\right) \\ &= P_{noise}\frac{G^Q - 1}{G - 1} \end{aligned} \tag{2.3}$$

Note that P_{noise} is the noise power that the amplifier itself contributes to the output. Assuming that the input signal comes from a passive device in thermal equilibrium, $N_{in} = \mathrm{kT}\Delta f$,

[1] Excluding the contribution of the source noise to the output.

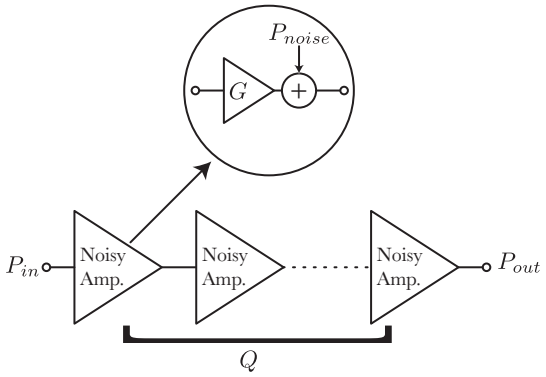


Figure 2.1: Chain of identical noisy amplifiers

where k is the Boltzmann constant, T is the absolute temperature, and Δf is the unit bandwidth, D can be calculated as

$$D = \frac{\frac{S_{in}}{\mathrm{kT}\Delta f}}{\frac{S_{in} \times G^Q}{P_{noise}\frac{G^Q-1}{G-1}}} \tag{2.4}$$

$$= \frac{P_{noise}}{\mathrm{kT}\Delta f}\frac{1-\frac{1}{G^Q}}{G-1} \tag{2.5}$$

Assuming that $G^Q \gg 1$, by cascading a large number of amplifiers or using a few amplifiers with high gain, we obtain a special case where

$$M = \frac{P_{noise}}{\mathrm{kT}\Delta f}\frac{1}{G-1} \tag{2.6}$$

where M represents the noise measure. Since the noise figure is $NF = 1 + \frac{P_{noise}}{\mathrm{kT}\Delta f G}$, M can be written as a function of the noise figure as

$$M = \frac{P_{noise}}{\mathrm{kT}\Delta f G}\frac{1}{1-\frac{1}{G}} \tag{2.7}$$

$$= \frac{NF-1}{1-\frac{1}{G}} \tag{2.8}$$

which is more common in the literature. It should be noted that:

- If the total power gain of the cascaded amplifiers is high enough ($G^M \gg 1$), the noise measure is an indicator of how much the SNR is degraded.

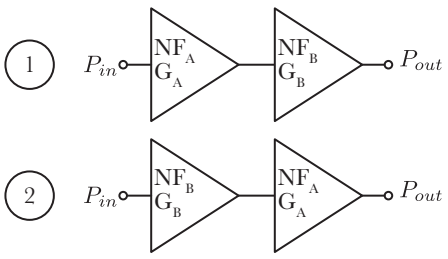


Figure 2.2: Two scenarios for cascading non-identical amplifiers

- When the power gain of a single stage is high enough ($G \gg 1$), the noise measure is $M \approx NF - 1$. Thus, as the frequency of the input signal approaches f_{max} of the active device, the noise measure becomes more critical.

The definition of power gain must be clarified here. Since kTΔf is the available noise power of the source, S_{in} must be considered as the (maximum available) power of the source. Therefore, it is reasonable to use the power gain and write the power of the output signal as:

$$S_{out} = S_{in,max} \times \left[\frac{P_{in,2}}{S_{in,max}} \times \frac{P_{in,3}}{P_{in,2}} \times ... \times \frac{P_L}{P_{in,M}} \right] \tag{2.9}$$

Since each $P_{in,j}$ appears once in the numerator and once in the denominator, they cancel each other. However, instead of canceling them, you can replace $P_{in,j}$ with $P_{out,max,j-1}$ and rewrite the previous equation as:

$$\begin{aligned} S_{out} &= S_{in,max} \times \frac{P_{out,max,1}}{S_{in,max}} \times ... \times \frac{P_L}{P_{out,max,M-1}} \\ &= S_{in,max} \times \frac{P_{out,max,1}}{S_{in,max}} \times ... \times \frac{P_{out,max,M}}{P_{out,max,M-1}} \times \frac{P_L}{P_{out,max,M}} \end{aligned}$$

Now, it is clear that the available power gain (G_{AP}) is the better choice when dealing with cascaded identical amplifiers because

$$S_{out} = S_{in} \times {G_{AP}}^M \times \left(1 - |\Gamma|^2\right) \tag{2.10}$$

$$\begin{aligned} N_{out} &= P_{noise,max} \left(1 + G_{AP} + ... + {G_{AP}}^{M-1}\right) \times \left(1 - |\Gamma|^2\right) \\ &= P_{noise,max} \frac{{G_{AP}}^M - 1}{G_{AP} - 1} \times \left(1 - |\Gamma|^2\right) \end{aligned} \tag{2.11}$$

where Γ is the output reflection coefficient of the last amplifier. Since there is $\frac{S_{out}}{N_{out}}$ in the definition of SNR degradation (Eq. 2.1), $(1 - |\Gamma|^2)$ cancels out. Therefore, in the rest of this chapter, the power gain of an amplifier is defined as its available power gain.

Now suppose that the cascaded amplifiers are not identical, as in Fig. 2.2. In scenario 1, amplifier "A" precedes amplifier "B", while in scenario 2, amplifier "B" is the first stage. Using the Friis formulas, the noise figure for each scenario can be calculated as

$$NF_1 = NF_A + \frac{NF_B - 1}{G_A} \tag{2.12}$$

$$NF_2 = NF_B + \frac{NF_A - 1}{G_B} \tag{2.13}$$

Comparing the two scenarios for the best noise figure (NF_1 and NF_2) gives the following:

$$\begin{aligned} NF_1 &\lesseqgtr NF_2 \\ NF_A + \frac{NF_B - 1}{G_A} &\lesseqgtr NF_B + \frac{NF_A - 1}{G_B} \\ NF_A - \frac{NF_A - 1}{G_B} &\lesseqgtr NF_B - \frac{NF_B - 1}{G_A} \\ NF_A - 1 - \frac{NF_A - 1}{G_B} &\lesseqgtr NF_B - 1 - \frac{NF_B - 1}{G_A} \\ (NF_A - 1)\left(1 - \frac{1}{G_B}\right) &\lesseqgtr (NF_B - 1)\left(1 - \frac{1}{G_A}\right) \end{aligned} \tag{2.14}$$

Assuming that the amplifiers have gain ($G_A > 1, G_B > 1$), the previous comparison in terms of noise measure can be written as

$$\begin{aligned} NF_1 &\lesseqgtr NF_2 \\ (NF_A - 1)\left(1 - \frac{1}{G_B}\right) &\lesseqgtr (NF_B - 1)\left(1 - \frac{1}{G_A}\right) \\ \frac{NF_A - 1}{1 - \frac{1}{G_A}} &\lesseqgtr \frac{NF_B - 1}{1 - \frac{1}{G_B}} \\ M_A &\lesseqgtr M_B \end{aligned} \tag{2.15}$$

The key observation is that it is better to start the amplification chain with the stage that has the lowest noise measure to achieve the minimum noise figure. The following section proves that the minimum noise measure is an invariant property of technology. It means that for any amplifier, if the gain of the amplifier increases, the noise figure must increase as consequently, as shown in Fig. 2.3.

2.2 Derivation of the Noise Measure

The minimum noise measure is calculated in [21], where the choice of circuit representation has led to unnecessarily complicated mathematical equations that are difficult to grasp

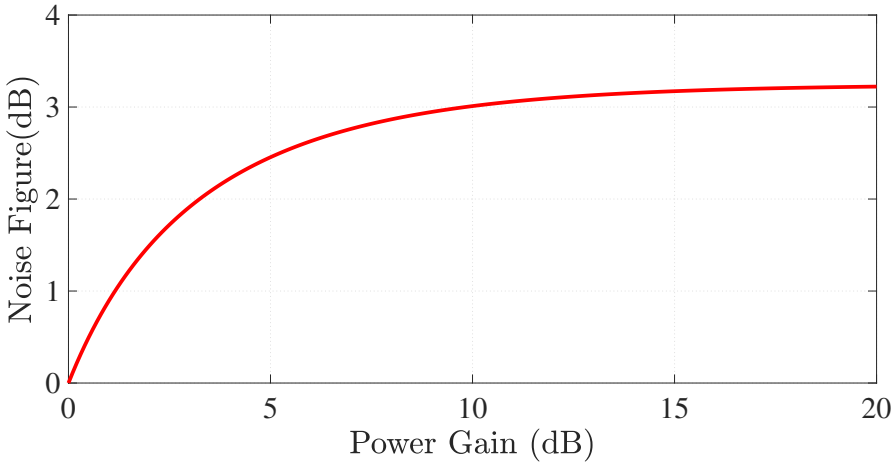


Figure 2.3: Noise figure vs. power gain

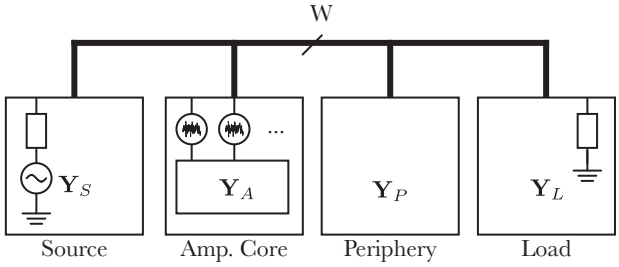


Figure 2.4: Y-parameter model of the circuit

intuitively. Here we provide a different circuit model (Fig. 2.4) in which all blocks are represented by their Y-parameter matrices ($\mathbf{Y}_{W\times W}$), which are connected to W voltage nodes (represented by $\mathbf{V}_{W\times 1}$) shared among them. The sub-indices S, A, P, and L represent the source, core amplifier, peripheral embedding network, and load, respectively. Without losing generality, the source and load ports are assumed to be connected between one of the W voltage nodes and the ground. The selection of the node for each port can be made using $\mathbf{u}_{W\times 1}$ vectors. For example, if the source port is connected to the first node

$$\mathbf{u}_S = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \dots \\ 0 \end{bmatrix}_{W\times 1} \tag{2.16}$$

In this case, the internal voltage source can be represented as

$$\mathbf{V}_S = v_S \mathbf{u}_S \tag{2.17}$$

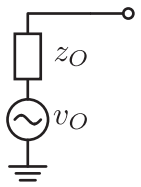


Figure 2.5: Thevenin equivalent circuit

where v_S is the physical internal voltage source. Similarly, $\mathbf{Y}_S$ can be defined as

$$\mathbf{Y}_S = y_S \mathbf{u}_S \mathbf{u}_S^H \tag{2.18}$$

where y_S is the physical source admittance, and $(.)^H$ represents the Hermitian transpose. Similarly, $\mathbf{Y}_L$ can be defined as.

$$\mathbf{Y}_L = y_L \mathbf{u}_L \mathbf{u}_L^H \tag{2.19}$$

Finally, the internal noise sources of the core amplifier are represented by W number of series noise voltages ($\mathbf{V}_N$) at each port of the amplifier. Writing the KCL equation for Fig. 2.4, we obtain

$$\mathbf{Y}_S (\mathbf{V} - \mathbf{V}_S) + \mathbf{Y}_L \mathbf{V} + \mathbf{Y}_P \mathbf{V} + \mathbf{Y}_A (\mathbf{V} - \mathbf{V}_N) = 0 \tag{2.20}$$

where $\mathbf{V}_{W\times 1}$ represents the voltage at each of the W nodes and can be calculated as

$$\mathbf{V} = [\mathbf{Y}_S + \mathbf{Y}_L + \mathbf{Y}_P + \mathbf{Y}_A]^{-1} [\mathbf{Y}_S \mathbf{V}_S + \mathbf{Y}_A \mathbf{V}_N] \tag{2.21}$$

To further simplify these equations, the effective $\mathbf{Y}_E$ matrix and the effective $\mathbf{I}_E$ matrix are defined as

$$\mathbf{Y}_E = \mathbf{Y}_S + \mathbf{Y}_L + \mathbf{Y}_P + \mathbf{Y}_A \tag{2.22}$$

$$\mathbf{I}_E = \mathbf{Y}_S \mathbf{V}_S + \mathbf{Y}_A \mathbf{V}_N \tag{2.23}$$

and thus $\mathbf{V} = {\mathbf{Y}_E}^{-1} \mathbf{I}_E$.

Before calculating the noise measure, we should clarify how to calculate the available power from the matrices defined earlier. If Fig. 2.5 represents the Thevenin equivalent circuit, then the available power of the load is

$$\begin{aligned} P_{O,max} &= \frac{1}{4} \frac{|v_O|^2}{\mathrm{Re}\{z_O\}} \\ &= \frac{1}{4} \frac{v_O {v_O}^*}{\frac{1}{2}(z_O + {z_O}^*)} \\ &= \frac{1}{2} \frac{v_O {v_O}^H}{z_O + {z_O}^H} \end{aligned} \tag{2.24}$$

where $(.)^*$ stands for the complex conjugate, corresponding to the Hermitian transpose operator when applied to a scalar number. The same equation can be used to find the available power of the source:

$$\begin{aligned} P_{S,max} &= \frac{1}{2}\frac{v_S {v_S}^H}{z_S + {z_S}^H} \\ &= \frac{1}{2}\frac{v_S {v_S}^H}{\frac{1}{y_S} + \frac{1}{{y_S}^H}} \\ &= \frac{1}{2}\frac{y_S v_S \times {v_S}^H {y_S}^H}{y_S + {y_S}^H} \end{aligned} \tag{2.25}$$

To apply the above equations, you must determine the Thevenin open-circuit voltage v_O and the output impedance z_O. To calculate v_O, the output port should be left open to calculate the output voltage [2]. It can be written as:

$$v_O = \mathbf{u}_L^H {\mathbf{Y}_{E,OC}}^{-1} \mathbf{I}_E \tag{2.26}$$

where $\mathbf{Y}_{E,OC}$ is defined as the effective y-parameter of the network when the output is open:

$$\mathbf{Y}_{E,OC} = \mathbf{Y}_S + \mathbf{Y}_P + \mathbf{Y}_A \tag{2.27}$$

To calculate the output impedance of the amplifier (z_O), you should calculate the output voltage response to a current test source at the output while all other independent sources are off. In matrix form, the following equation should be solved to find the voltage vector $\mathbf{V}$:

$$\mathbf{Y}_{E,OC}\mathbf{V} = i_t \mathbf{u}_L \tag{2.28}$$

where i_t is the test current at the output port. Once $\mathbf{V}$ is calculated, the output impedance can be calculated by dividing the output voltage by the test current source. In matrix form:

$$\mathbf{V} = {\mathbf{Y}_{E,OC}}^{-1} i_t \mathbf{u}_L \tag{2.29}$$

and therefore

$$\begin{aligned} z_O &= \frac{1}{i_t}\mathbf{u}_L^H {\mathbf{Y}_{E,OC}}^{-1} i_t \mathbf{u}_L \\ &= \mathbf{u}_L^H {\mathbf{Y}_{E,OC}}^{-1} \mathbf{u}_L \end{aligned} \tag{2.30}$$

[2] After calculating the voltage vector $\mathbf{V}$, it should be multiplied by $\mathbf{u}_L^H$ to extract the (scalar) output voltage from it

Using Eq. 2.24, the available output power can be written as

$$\begin{aligned} P_{O,max} &= \frac{1}{2}\frac{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\mathbf{I}_E \times \{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\mathbf{I}_E\}^H}{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\mathbf{u}_L + \{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\mathbf{u}_L\}^H} \\ &= \frac{1}{2}\frac{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\mathbf{I}_E \times \mathbf{I}_E{}^H \mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L}{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\mathbf{u}_L + \mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L} \\ &= \frac{1}{2}\frac{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{I}_E \mathbf{I}_E{}^H\right)\mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L}{\mathbf{u}_L^H \left(\mathbf{Y}_{E,OC}{}^{-1} + \mathbf{Y}_{E,OC}{}^{-1^H}\right)\mathbf{u}_L} \\ &= \frac{1}{2}\frac{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{I}_E \mathbf{I}_E{}^H\right)\mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L}{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_{E,OC} + \mathbf{Y}_{E,OC}{}^H\right)\mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L} \end{aligned} \tag{2.31}$$

To calculate the noise measure, noise and gain must be calculated from Eq. 2.6. The available power gain can be calculated by substituting $\mathbf{I}_E = \mathbf{Y}_S\mathbf{V}_S$ into Eq. 2.31 and normalizing with the available input power from Eq. 2.25.

$$\begin{aligned} G =& \frac{\frac{1}{2}\frac{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_S\mathbf{V}_S(\mathbf{Y}_S\mathbf{V}_S)^H\right)\mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L}{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_{E,OC}+\mathbf{Y}_{E,OC}{}^H\right)\mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L}}{\frac{1}{2}\frac{y_S v_S \times v_S{}^H y_S{}^H}{y_S + y_S{}^H}} \\ =& \frac{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_S\mathbf{V}_S\mathbf{V}_S{}^H\mathbf{Y}_S{}^H\right)\mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L}{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_{E,OC} + \mathbf{Y}_{E,OC}{}^H\right)\mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L} \times \frac{y_S + y_S{}^H}{y_S v_S \times v_S{}^H y_S{}^H} \end{aligned} \tag{2.32}$$

Note that all components in the numerator and denominator of the second fraction are scalar and can be freely shifted in the multiplication chain. Using Eq. 2.17 and Eq. 2.18, the available gain can be written as in Eq. 2.33.

$$\begin{aligned} G &= \frac{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\frac{y_S + y_S{}^H}{y_S v_S \times v_S{}^H y_S{}^H}\mathbf{Y}_S\mathbf{V}_S\mathbf{V}_S{}^H\mathbf{Y}_S{}^H\right)\mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L}{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_{E,OC} + \mathbf{Y}_{E,OC}{}^H\right)\mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L} \\ &= \frac{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\left(y_S + y_S{}^H\right)\mathbf{u}_S\mathbf{u}_S^H\mathbf{u}_S\mathbf{u}_S^H\mathbf{u}_S\mathbf{u}_S^H\right)\mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L}{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_{E,OC} + \mathbf{Y}_{E,OC}{}^H\right)\mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L} \\ &= \frac{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\left(y_S + y_S{}^H\right)\mathbf{u}_S\mathbf{u}_S^H\right)\mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L}{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_{E,OC} + \mathbf{Y}_{E,OC}{}^H\right)\mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L} \\ &= \frac{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_S + \mathbf{Y}_S{}^H\right)\mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L}{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_{E,OC} + \mathbf{Y}_{E,OC}{}^H\right)\mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L} \end{aligned} \tag{2.33}$$

Similarly, the noise power can be calculated by substituting $\mathbf{I}_E = \mathbf{Y}_A\mathbf{V}_N$ into Eq. 2.31.

$$P_{noise} = \frac{1}{2}\frac{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_A\mathbf{V}_N\mathbf{V}_N{}^H\mathbf{Y}_A{}^H\right)\mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L}{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_{E,OC} + \mathbf{Y}_{E,OC}{}^H\right)\mathbf{Y}_{E,OC}{}^{-1^H}\mathbf{u}_L} \tag{2.34}$$

Using Eq. 2.6, the noise measure can be written as in Eq. 2.35.

$$\begin{aligned}
M &= \frac{1}{\mathrm{kT}\Delta f}\frac{\frac{1}{2}\frac{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_A \mathbf{V}_N \mathbf{V}_N{}^H \mathbf{Y}_A{}^H\right)\mathbf{Y}_{E,OC}{}^{-1H}\mathbf{u}_L}{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_{E,OC}+\mathbf{Y}_{E,OC}{}^H\right)\mathbf{Y}_{E,OC}{}^{-1H}\mathbf{u}_L}}{\frac{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_S+\mathbf{Y}_S{}^H\right)\mathbf{Y}_{E,OC}{}^{-1H}\mathbf{u}_L}{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_{E,OC}+\mathbf{Y}_{E,OC}{}^H\right)\mathbf{Y}_{E,OC}{}^{-1H}\mathbf{u}_L}} - 1 \\
&= \frac{1}{2\mathrm{kT}\Delta f}\frac{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_A \mathbf{V}_N \mathbf{V}_N{}^H \mathbf{Y}_A{}^H\right)\mathbf{Y}_{E,OC}{}^{-1H}\mathbf{u}_L}{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_S+\mathbf{Y}_S{}^H\right)\mathbf{Y}_{E,OC}{}^{-1H}\mathbf{u}_L - \mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_{E,OC}+\mathbf{Y}_{E,OC}{}^H\right)\mathbf{Y}_{E,OC}{}^{-1H}\mathbf{u}_L} \\
&= \frac{-1}{2\mathrm{kT}\Delta f}\frac{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_A \mathbf{V}_N \mathbf{V}_N{}^H \mathbf{Y}_A{}^H\right)\mathbf{Y}_{E,OC}{}^{-1H}\mathbf{u}_L}{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_{E,OC}+\mathbf{Y}_{E,OC}{}^H-\mathbf{Y}_S-\mathbf{Y}_S{}^H\right)\mathbf{Y}_{E,OC}{}^{-1H}\mathbf{u}_L} \\
&= \frac{-1}{2\mathrm{kT}\Delta f}\frac{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_A \mathbf{V}_N \mathbf{V}_N{}^H \mathbf{Y}_A{}^H\right)\mathbf{Y}_{E,OC}{}^{-1H}\mathbf{u}_L}{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_A+\mathbf{Y}_A{}^H+\mathbf{Y}_P+\mathbf{Y}_P{}^H\right)\mathbf{Y}_{E,OC}{}^{-1H}\mathbf{u}_L}
\end{aligned} \tag{2.35}$$

To gain intuition over the previous equation, we assume that the peripheral network neither absorbs nor generates energy. For example, the peripheral network may consist of passive, loss-less components [3]. In this case, the total active power in this block should be zero

$$\begin{aligned}
P_{loss-less} &= 0 \\
&= \mathrm{Re}\{\mathbf{V}^H \times \mathbf{Y}_P \mathbf{V}\} \\
&= \frac{1}{2}\left[\mathbf{V}^H \times \mathbf{Y}_P \mathbf{V} + \left(\mathbf{V}^H \times \mathbf{Y}_P \mathbf{V}\right)^H\right] \\
&= \frac{1}{2}\left[\mathbf{V}^H \times \mathbf{Y}_P \mathbf{V} + \mathbf{V}^H \mathbf{Y}_P{}^H \times \mathbf{V}\right] \\
&= \frac{1}{2}\mathbf{V}^H\left[\mathbf{Y}_P + \mathbf{Y}_P{}^H\right]\mathbf{V}
\end{aligned} \tag{2.36}$$

Since this equation must hold for all possible $\mathbf{V}$ vectors, it can be concluded that

$$\mathbf{Y}_P + \mathbf{Y}_P{}^H\big|_{\mathbf{Y}_P:loss-less} = 0 \tag{2.37}$$

Therefore, under the assumption of a passive loss-less peripheral network, Eq. 2.35 can be simplified to

$$M = \frac{-1}{2\mathrm{kT}\Delta f} \times \frac{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_A \mathbf{V}_N \mathbf{V}_N{}^H \mathbf{Y}_A{}^H\right)\mathbf{Y}_{E,OC}{}^{-1H}\mathbf{u}_L}{\mathbf{u}_L^H \mathbf{Y}_{E,OC}{}^{-1}\left(\mathbf{Y}_A+\mathbf{Y}_A{}^H\right)\mathbf{Y}_{E,OC}{}^{-1H}\mathbf{u}_L} \tag{2.38}$$

The minimum noise measure is desired here since it sets the lower bound of the SNR in a low-noise amplification chain. To simplify the calculations, new symbols are defined as

[3] Note that network reciprocity is not used here and the only assumption is that the power flow is zero

follows:

$$\mathbf{x}_{W\times 1} = \mathbf{Y}_{E,OC}{}^{-1H}\mathbf{u}_L \tag{2.39}$$

$$\mathbf{A}_{W\times W} = \mathbf{Y}_A\mathbf{V}_N\mathbf{V}_N{}^H\mathbf{Y}_A{}^H \tag{2.40}$$

$$\mathbf{B}_{W\times W} = -2\mathrm{kT}\Delta f\left(\mathbf{Y}_A + \mathbf{Y}_A{}^H\right) \tag{2.41}$$

Note that while $\mathbf{A}$ and $\mathbf{B}$ [4] are fixed by the available active device, the vector $\mathbf{x}$ can be modified by the proper choice of load port and peripheral network. Therefore, the noise measure $M_{\mathbf{x}}$[5] is only a function of the vector $\mathbf{x}$

$$M_{\mathbf{x}} = \frac{\mathbf{x}^H\mathbf{A}\mathbf{x}}{\mathbf{x}^H\mathbf{B}\mathbf{x}} \tag{2.42}$$

Therefore, $M_{\mathbf{x}}$ should be minimized under the constraint $g(\mathbf{x}) = \mathbf{x}^H\mathbf{A}\mathbf{x} - M_{\mathbf{x}}\mathbf{x}^H\mathbf{B}\mathbf{x} = 0$ or

$$g(\mathbf{x}) = \mathbf{x}^H\left(\mathbf{A} - M_{\mathbf{x}}\mathbf{B}\right)\mathbf{x} = 0 \tag{2.43}$$

Since $g(\mathbf{x}) = 0$ is constant, its derivative with respect to the real and imaginary parts of each component x_j of the vector $\mathbf{x}$ should be zero. With respect to the real parts of each component $x_{j,re}$

$$\begin{aligned}\frac{\partial g(\mathbf{x})}{\partial x_{j,re}} =& 0\\ =& \left(\frac{\partial \mathbf{x}}{\partial x_{j,re}}\right)^H\left(\mathbf{A} - M_{\mathbf{x}}\mathbf{B}\right)\mathbf{x}\\ &+ \mathbf{x}^H\left(-\frac{\partial M_{\mathbf{x}}}{\partial x_{j,re}}\mathbf{B}\right)\mathbf{x}\\ &+ \mathbf{x}^H\left(\mathbf{A} - M_{\mathbf{x}}\mathbf{B}\right)\frac{\partial \mathbf{x}}{\partial x_{j,re}}\end{aligned} \tag{2.44}$$

and with respect to the imaginary parts of each component $x_{j,im}$

$$\begin{aligned}\frac{\partial g(\mathbf{x})}{\partial x_{j,im}} =& 0\\ =& \left(\frac{\partial \mathbf{x}}{\partial x_{j,im}}\right)^H\left(\mathbf{A} - M_{\mathbf{x}}\mathbf{B}\right)\mathbf{x}\\ &+ \mathbf{x}^H\left(-\frac{\partial M_{\mathbf{x}}}{\partial x_{j,im}}\mathbf{B}\right)\mathbf{x}\\ &+ \mathbf{x}^H\left(\mathbf{A} - M_{\mathbf{x}}\mathbf{B}\right)\frac{\partial \mathbf{x}}{\partial x_{j,im}}\end{aligned} \tag{2.45}$$

[4] $\mathbf{A}$ and $\mathbf{B}$ are both Hermitian. Moreover, $\mathbf{A}$ is also positive (semi)-definite.

[5] The subindex $\mathbf{x}$ is used to emphasize that M is a function of $\mathbf{x}$.

For the optimal noise measure (λ), $\frac{\partial M_{\mathbf{x}}}{\partial x_{j,re}} = \frac{\partial M_{\mathbf{x}}}{\partial x_{j,im}} = 0$. Moreover, $\frac{\partial \mathbf{x}}{\partial x_{j,im}} = \mathrm{i}\frac{\partial \mathbf{x}}{\partial x_{j,re}}$[6]. To satisfy the previous equations,

$$(\mathbf{A} - \lambda\mathbf{B})\,\mathbf{x} = \mathbf{B}\left(\mathbf{B}^{-1}\mathbf{A} - \lambda\mathbf{I}\right)\mathbf{x} = 0 \tag{2.46}$$

In other words, all local optima of the noise measure (λ) are eigenvalues of the characteristic noise matrix ($\mathbf{N}$), defined as

$$\begin{aligned}\mathbf{N} &= \mathbf{B}^{-1}\mathbf{A} \\ &= \frac{-1}{2\mathrm{kT}\Delta f}\left(\mathbf{Y}_A + \mathbf{Y}_A{}^H\right)^{-1}\mathbf{Y}_A\mathbf{V}_N\mathbf{V}_N{}^H\mathbf{Y}_A{}^H \end{aligned}\tag{2.47}$$

and the minimum noise measure is equal to the minimum eigenvalue (λ_{min}) of this matrix. Since the noise is stochastic, $\mathbf{V}_N\mathbf{V}_N{}^H$ should be replaced by the noise correlation matrix

$$\begin{aligned}\mathbf{N} &= \mathbf{B}^{-1}\mathbf{A} \\ &= \frac{-1}{2\mathrm{kT}\Delta f}\left(\mathbf{Y}_A + \mathbf{Y}_A{}^H\right)^{-1}\mathbf{Y}_A\overline{\mathbf{V}_N\mathbf{V}_N{}^H}\mathbf{Y}_A{}^H &(2.48)\\ &= \frac{-1}{2\mathrm{kT}\Delta f}\left(\mathbf{Y}_A + \mathbf{Y}_A{}^H\right)^{-1}\overline{\mathbf{I}_N\mathbf{I}_N{}^H} &(2.49)\end{aligned}$$

where $\mathbf{I}_N = \mathbf{Y}_A\mathbf{V}_N$. Similar results are obtained in [22]. The important conclusion is that using a passive loss-less embedding network does not change the minimum achievable noise measure. To achieve the minimum noise measure, circuits must be designed such that $\mathbf{x}$ is the (α scaled) eigenvector $\mathbf{e}_{\lambda_{min}}$ of $\mathbf{N}$ corresponding to the minimum eigenvalue λ_{min}, which implies

$$\mathbf{Y}_{E,OC}{}^{-1H}\mathbf{u}_L = \alpha\mathbf{e}_{\lambda_{min}} \tag{2.50}$$

Assuming that $\mathbf{u}_L$ and $\mathbf{u}_S$ are real vectors, the previous equation can be simplified as

$$\begin{aligned}\beta\mathbf{u}_L &= \mathbf{Y}_{E,OC}{}^T\mathbf{e}^*_{\lambda_{min}} \\ &= \left(\mathbf{Y}_S + \mathbf{Y}_P + \mathbf{Y}_A\right)^T\mathbf{e}^*_{\lambda_{min}}\end{aligned}\tag{2.51}$$

Where $\beta = \frac{1}{\alpha^*}$ is used. Therefore, the peripheral network and the source impedance must be designed such that

$$y_S\mathbf{u}_S\mathbf{u}_S^H\mathbf{e}^*_{\lambda_{min}} = \beta\mathbf{u}_L - \left(\mathbf{Y}_P + \mathbf{Y}_A\right)^T\mathbf{e}^*_{\lambda_{min}} \tag{2.52}$$

Since y_S and β are two complex independent variables, the above equation can always be satisfied in a two-port network, resulting in a minimum noise figure. Thus, the minimum noise figure is just a function of the inherent characteristics of the active device.

[6] Alternatively, the Cauchy-Riemann equations can be used

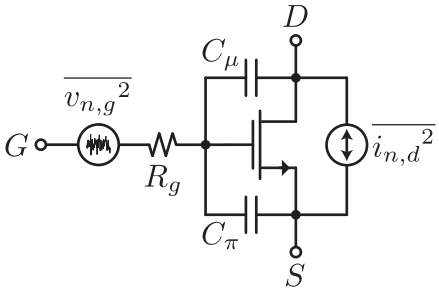


Figure 2.6: CMOS transistor parasitic model.

2.3 Examples

CMOS Noise Measure

To calculate the minimum noise measure, a CMOS transistor is modeled as shown in Fig. 2.6[7]. $\mathbf{Y}_A$ can be written as

$$\mathbf{Y}_A = \begin{bmatrix} \dfrac{(C_\pi + C_\mu)S}{R_g(C_\pi + C_\mu)S + 1} & \dfrac{-C_\mu S}{R_g(C_\pi + C_\mu)S + 1} \\ \dfrac{g_m - C_\mu S}{R_g(C_\pi + C_\mu)S + 1} & \dfrac{(R_g C_\pi S + g_m R_g + 1)\, C_\mu S}{R_g(C_\pi + C_\mu)S + 1} \end{bmatrix} \tag{2.53}$$

which satisfies KCL equation of

$$\begin{bmatrix} i_G \\ i_D \end{bmatrix} = \mathbf{Y}_A \begin{bmatrix} v_G \\ v_D \end{bmatrix} \tag{2.54}$$

The effective current noise source can be written as

$$\mathbf{I}_N = \begin{bmatrix} 0 \\ i_{n,d} \end{bmatrix} + \mathbf{Y}_A \begin{bmatrix} v_{n,g} \\ 0 \end{bmatrix} \tag{2.55}$$

Note that two noise sources shown in Fig. 2.6 are independent and have a power spectral density of

$$\overline{i_{n,d}{}^2} = 4\mathrm{kT}\gamma g_m \Delta f \tag{2.56}$$

$$\overline{v_{n,g}{}^2} = 4\mathrm{kT} R_g \Delta f \tag{2.57}$$

where the channel-induced gate current noise is ignored [23]. Assuming no correlation between noise sources $< i_{n,d}, v_{n,g} >= 0$,

$$\overline{\mathbf{I}_N \mathbf{I}_N{}^H} = \begin{bmatrix} 0 & 0 \\ 0 & \overline{i_{n,d}{}^2} \end{bmatrix} + \mathbf{Y}_A \begin{bmatrix} \overline{v_{n,g}{}^2} & 0 \\ 0 & 0 \end{bmatrix} \mathbf{Y}_A{}^H \tag{2.58}$$

[7]It should be noted that without loss of generality, all purely imaginary parasitic elements at the gain/source/drain nodes can be considered as part of the loss-less peripheral network

Consequently, the characteristic noise matrix can be calculated using Eq. 2.47. This matrix has two eigenvalues:

$$\lambda = \frac{1}{1-\frac{1}{U}}\left[2\kappa + \frac{1}{2U} \pm \sqrt{4\kappa + \left(2\kappa - \frac{1}{2U}\right)^2}\right] \tag{2.59}$$

where U is Mason's unilateral power gain, which can be calculated as [24][25]

$$U = \frac{1}{4R_g\left(g_{ds} + g_m\frac{C_\mu}{C_\mu + C_\pi}\right)}\left(\frac{\omega_T}{\omega}\right)^2 \tag{2.60}$$

and

$$\kappa = \gamma g_m R_g \left(\frac{\omega}{\omega_T}\right)^2 \tag{2.61}$$

$$\omega_T = \frac{g_m}{C_\mu + C_\pi} \tag{2.62}$$

Since a positive power gain corresponds to a positive noise measure, only the positive eigenvalue is acceptable. Therefore, for a CMOS amplifier, the minimum achievable noise measure is

$$M_{min} = \frac{1}{1-\frac{1}{U}}\left[2\kappa + \frac{1}{2U} + \sqrt{4\kappa + \left(2\kappa - \frac{1}{2U}\right)^2}\right] \tag{2.63}$$

To gain insight into the behavior of the noise measure as a function of frequency, the above equation should be simplified by making reasonable assumptions. First, note that while both U and κ are frequency-dependent, when

$$\frac{g_{ds}}{g_m} + \frac{C_\mu}{C_\mu + C_\pi} \ll \gamma \tag{2.64}$$

then it can be concluded that

$$\frac{1}{2U} \ll 2\kappa \tag{2.65}$$

which is independent of the operating frequency and results in

$$M_{min} \approx \frac{1}{1-\frac{1}{U}}\left[2\kappa + \sqrt{4\kappa + 4\kappa^2}\right] \tag{2.66}$$

At low frequencies where $\kappa < 1$,

$$M_{min} \approx \frac{2\sqrt{\kappa}}{1-\frac{1}{U}}\left(1 + \sqrt{\kappa} + \frac{\kappa}{2}\right) \tag{2.67}$$

and it can be observed that the minimum noise measure increases as a linear function of frequency

$$M_{min}|_{\omega \ll \omega_T} \approx \sqrt{4\gamma g_m R_g}\frac{\omega}{\omega_T} \tag{2.68}$$

which shows a similar trend as the minimum noise figure of the transistor. However, as the operating frequency approaches f_{max} of the device where $U = 1$, the noise measure approaches infinity. Fig. 2.7 shows a comparison between the simulation of a commercial CMOS 28nm PDK (post layout extraction) [8] vs. the calculation results of Eq. 2.63 with the parameters from Table. 2.1[9]. Even at the frequency of $f = \frac{f_{max}}{2}$, the error of Eq. 2.68 is less than 50%. As a rule of thumb, this frequency should be used to evaluate the feasibility of implementing low-noise amplifiers for any technology.

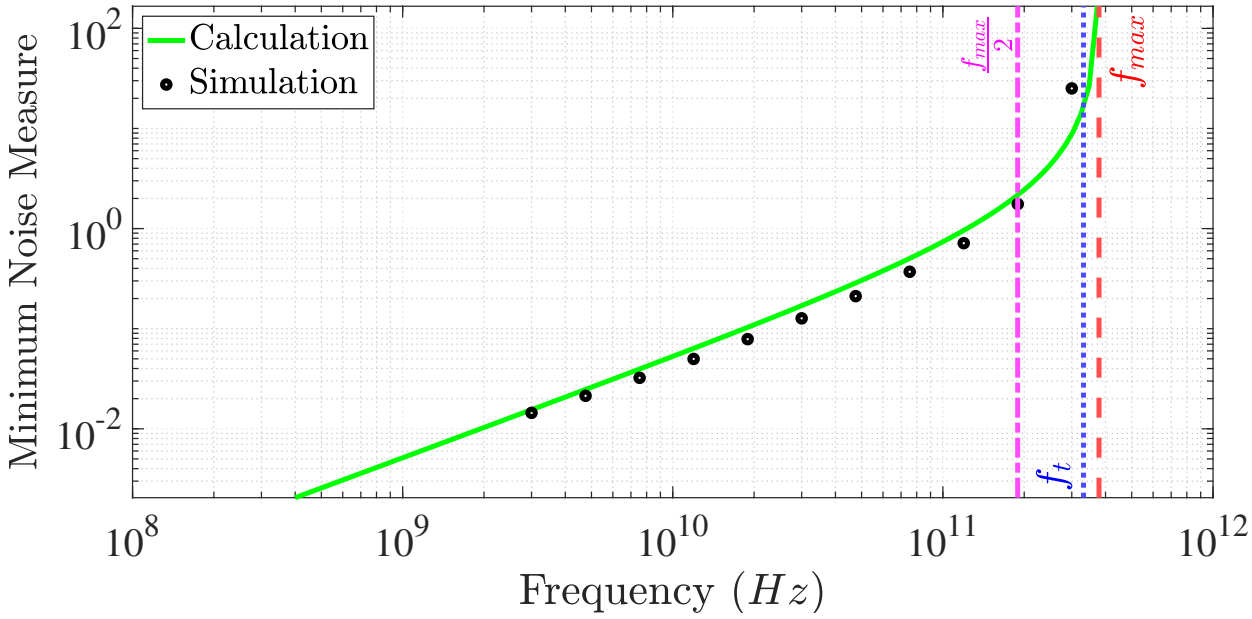


Figure 2.7: Noise measure vs. frequency

Multiple Active Devices

Since a single device has a minimum noise measure, one might think that using multiple active devices would improve performance. For example, in a noise-canceling LNA, an auxiliary amplifier helps reduce the main amplifier's noise. In this section, we present a simple case with two amplifiers to show that the minimum noise measure of a single device also dictates the minimum noise measure of any combination of multiple devices. Assuming that the noise

[8] The simulation method is explained in the appendix.

[9] Note that the parameters from Table. 2.1 are only used to show the trends based on the calculations and are not extracted from the PDK.

Parameter	Vale
γ	2
g_m	$25mS$
R_g	14.4Ω
C_μ	$4fF$
C_π	$8fF$
f_t	$315GHz$
f_{max}	$378GHz$

Table 2.1: Parameter values used for calculations

($\mathbf{I}_{N1}$,$\mathbf{I}_{N2}$) of two amplifiers ($\mathbf{Y}_{A1}$,$\mathbf{Y}_{A2}$) is uncorrelated, the characteristic noise matrix can be written as in Eq. 2.69

$$\begin{aligned}
\mathbf{N} &= \frac{-1}{2\mathrm{kT}\Delta f}\left(\mathbf{Y}_A+\mathbf{Y}_A{}^H\right)^{-1}\overline{\mathbf{I}_N\mathbf{I}_N{}^H} \\
&= \frac{-1}{2\mathrm{kT}\Delta f}\begin{bmatrix}\mathbf{Y}_{A1}+\mathbf{Y}_{A1}{}^H & \mathbf{0}\\ \mathbf{0} & \mathbf{Y}_{A2}+\mathbf{Y}_{A2}{}^H\end{bmatrix}^{-1}\begin{bmatrix}\overline{\mathbf{I}_{N1}\mathbf{I}_{N1}{}^H} & \mathbf{0}\\ \mathbf{0} & \overline{\mathbf{I}_{N2}\mathbf{I}_{N2}{}^H}\end{bmatrix} \\
&= \frac{-1}{2\mathrm{kT}\Delta f}\begin{bmatrix}\left(\mathbf{Y}_{A1}+\mathbf{Y}_{A1}{}^H\right)^{-1} & \mathbf{0}\\ \mathbf{0} & \left(\mathbf{Y}_{A2}+\mathbf{Y}_{A2}{}^H\right)^{-1}\end{bmatrix}\begin{bmatrix}\overline{\mathbf{I}_{N1}\mathbf{I}_{N1}{}^H} & \mathbf{0}\\ \mathbf{0} & \overline{\mathbf{I}_{N2}\mathbf{I}_{N2}{}^H}\end{bmatrix} \\
&= \frac{-1}{2\mathrm{kT}\Delta f}\begin{bmatrix}\left(\mathbf{Y}_{A1}+\mathbf{Y}_{A1}{}^H\right)^{-1}\overline{\mathbf{I}_{N1}\mathbf{I}_{N1}{}^H} & \mathbf{0}\\ \mathbf{0} & \left(\mathbf{Y}_{A2}+\mathbf{Y}_{A2}{}^H\right)^{-1}\overline{\mathbf{I}_{N2}\mathbf{I}_{N2}{}^H}\end{bmatrix}
\end{aligned} \tag{2.69}$$

Since the eigenvalues of a block-diagonal matrix are the combination of the eigenvalues of the original sub-matrices, the minimum noise measure is equal to the minimum noise measures of the two amplifiers. Therefore, noise-canceling topologies cannot improve the minimum noise measure.

2.4 Design of Low-Noise CS Amplifiers with Single Feedback Component

In this part, we design a simple common-source stage for the minimum noise measure. As mentioned earlier, it is always possible to achieve the minimum noise measure in a two-port network. However, the correct value of the source impedance may be far from the matching condition. Here, a passive, reactive feedback component from the drain to the gate of the transistor is used to improve the input reflection of the LNA, as shown in Fig. 2.8. The following procedure is used in the simulation:

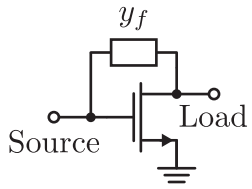


Figure 2.8: Using a feedback component to improve the input reflection with minimum noise measure

1. $\mathbf{Y}_A$ and $\mathbf{e}_{\lambda_{min}}$ are extracted to be used in Eq. 2.52.
2. Given the admittance y_f of the feedback component,

$$\mathbf{Y}_P = \begin{bmatrix} y_f & -y_f \\ -y_f & y_f \end{bmatrix} \tag{2.70}$$

the optimal source impedance can be calculated from Eq. 2.52 as

$$y_S e_1^* = -(y_f + y_{A_{11}})e_1^* - (-y_f + y_{A_{21}})e_2^* \tag{2.71}$$

where $\mathbf{e}_{\lambda_{min}} = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$. Therefore, the minimum noise measure is guaranteed when

$$y_{S_{Opt}} = -(y_f + y_{A_{11}}) + (y_f - y_{A_{21}})\frac{e_2^*}{e_1^*} \tag{2.72}$$

3. While sweeping the feedback admittance, the maximum available power gain (G_{max}) is measured and compared to the available power gain (G_a) for the source impedance calculated for the minimum noise measure ($y_{S_{Opt}}$). The admittance which corresponds to the smallest difference between two gain metrics (G_{max} and G_a) is the optimal feedback admittance for a low-noise CS stage to satisfy the minimum noise measure of the technology while minimizing input reflection.
4. When multiple stages are cascaded, the output matching network should be designed so that the effective source impedance seen by each stage is equal to the optimal source impedance ($y_{S_{Opt}}$). Otherwise, the output matching network should be designed to achieve a conjugate match at the output when a single stage is used. The key observation here is that the minimum noise figure is guaranteed as long as the source impedance of $y_{S_{Opt}}$ is provided to each stage.

Fig. 2.9 shows the optimum feedback admittance and various power gain factors at 190GHz for a 28nm CMOS transistor after parasitic extraction. It should be noted that a neutralized device does not necessarily have the minimum noise measure. Moreover, designers can meet the minimum noise measure of technology while achieving a power gain much higher than the Mason's unilateral power gain [26].

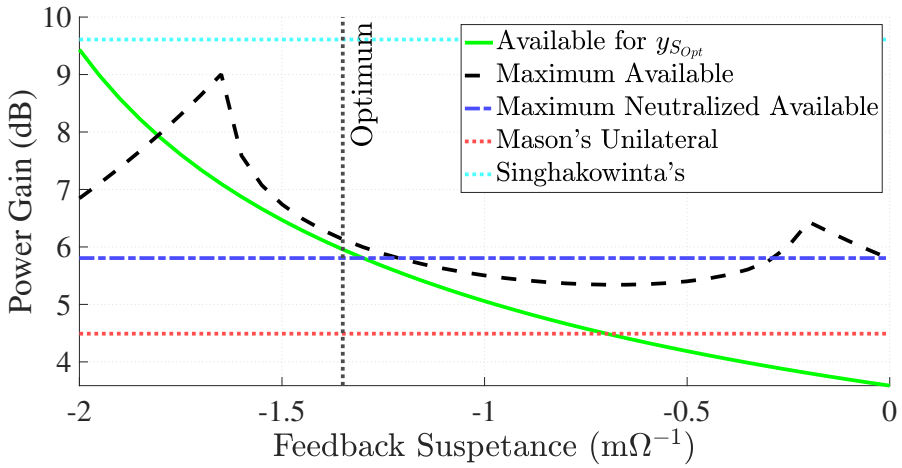
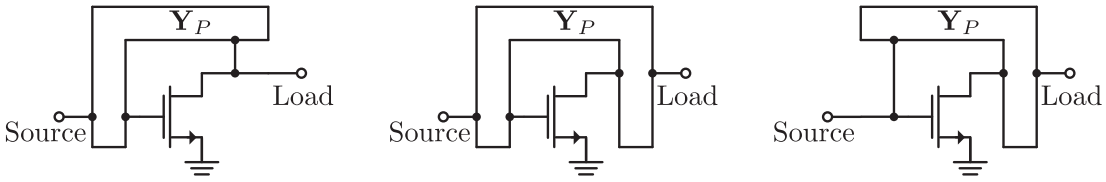


Figure 2.9: Power gain vs. feedback admittance at 190GHz.

2.5 Design of Low-Noise CS Amplifiers with General Peripheral Network



(a) Source-side (b) Double-side (c) Load-side

Figure 2.10: Different peripheral networks

As we have already shown, a feedback component at the gate-drain ports can only improve the input reflection to a limited extent while achieving the minimum noise measure. In this section, we consider general passive, reactive, and reciprocal peripheral networks, as shown in Fig. 2.10, to investigate whether the minimum noise measure can be achieved with a simultaneous conjugate match at the input and output. Starting from a general double-side tuning (Fig. 2.10b)

$$y_S \mathbf{u}_S \mathbf{u}_S^H \mathbf{e}'^*_{\lambda_{min}} = \beta \mathbf{u}_L - \left(\mathbf{Y}_P + \mathbf{Y}'_A\right)^T \mathbf{e}'^*_{\lambda_{min}} \tag{2.73}$$

where

$$\mathbf{Y}'_A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & y_{A_{11}} & y_{A_{12}} & 0 \\ 0 & y_{A_{21}} & y_{A_{22}} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.74}$$

$$\mathbf{e'}_{\lambda_{min}} = \begin{bmatrix} 0 \\ e_1 \\ e_2 \\ 0 \end{bmatrix} \tag{2.75}$$

$$\mathbf{Y'}_P = \begin{bmatrix} y_{P_{11}} & y_{P_{12}} & y_{P_{13}} & y_{P_{14}} \\ y_{P_{12}} & y_{P_{22}} & y_{P_{23}} & y_{P_{24}} \\ y_{P_{13}} & y_{P_{23}} & y_{P_{33}} & y_{P_{34}} \\ y_{P_{14}} & y_{P_{24}} & y_{P_{34}} & y_{P_{44}} \end{bmatrix} \tag{2.76}$$

where an order of source, gate, drain, and load is used for the port indices. Note that under the assumption that the components of $y_P \neq \infty$ are finite, the minimum noise measure can be obtained when

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \beta \end{bmatrix} - \begin{bmatrix} 0 & y_{P_{12}} & y_{P_{13}} & 0 \\ 0 & y_{A_{11}} + y_{P_{22}} & y_{A_{21}} + y_{P_{23}} & 0 \\ 0 & y_{A_{12}} + y_{P_{23}} & y_{A_{22}} + y_{P_{33}} & 0 \\ 0 & y_{P_{24}} & y_{P_{34}} & 0 \end{bmatrix} \begin{bmatrix} 0 \\ e_1^* \\ e_2^* \\ 0 \end{bmatrix} \tag{2.77}$$

is satisfied. Unfortunately, with three purely imaginary variables $y_{P_{22}}$, $y_{P_{23}}$ and $y_{P_{33}}$, the four equations resulting from the real and imaginary parts of the second and third lines cannot be satisfied.

For a source-side tuning (Fig. 2.10a), the following equations are defined:

$$\mathbf{Y'}_A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & y_{A_{11}} & y_{A_{12}} \\ 0 & y_{A_{21}} & y_{A_{22}} \end{bmatrix} \tag{2.78}$$

$$\mathbf{e'}_{\lambda_{min}} = \begin{bmatrix} 0 \\ e_1 \\ e_2 \end{bmatrix} \tag{2.79}$$

$$\mathbf{Y'}_P = \begin{bmatrix} y_{P_{11}} & y_{P_{12}} & y_{P_{13}} \\ y_{P_{12}} & y_{P_{22}} & y_{P_{23}} \\ y_{P_{13}} & y_{P_{23}} & y_{P_{33}} \end{bmatrix} \tag{2.80}$$

where an order of source, gate, and load is used for the port indices. The minimum noise measure is reached when

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \beta \end{bmatrix} - \begin{bmatrix} 0 & y_{P_{12}} & y_{P_{13}} \\ 0 & y_{A_{11}} + y_{P_{22}} & y_{A_{21}} + y_{P_{23}} \\ 0 & y_{A_{12}} + y_{P_{23}} & y_{A_{22}} + y_{P_{33}} \end{bmatrix} \begin{bmatrix} 0 \\ e_1^* \\ e_2^* \end{bmatrix} \tag{2.81}$$

is satisfied. While the previous one no longer exists, the first line with two purely imaginary $y_{P_{12}}$ and $y_{P_{13}}$ can only be satisfied if only $\frac{e_1}{e_2}$ is purely real, a condition that is not usually satisfied.

If you use load-side tuning (Fig. 2.10c), you get the following equations:

$$\mathbf{Y'}_A = \begin{bmatrix} 0 & 0 & 0 \\ y_{A_{11}} & y_{A_{12}} & 0 \\ y_{A_{21}} & y_{A_{22}} & 0 \end{bmatrix} \tag{2.82}$$

$$\mathbf{e'}_{\lambda_{min}} = \begin{bmatrix} e_1 \\ e_2 \\ 0 \end{bmatrix} \tag{2.83}$$

$$\mathbf{Y'}_P = \begin{bmatrix} y_{P_{11}} & y_{P_{12}} & y_{P_{13}} \\ y_{P_{12}} & y_{P_{22}} & y_{P_{23}} \\ y_{P_{13}} & y_{P_{23}} & y_{P_{33}} \end{bmatrix} \tag{2.84}$$

where an order of source, drain, and load is used for the port indices. The minimum noise measure is obtained when

$$\begin{bmatrix} y_S e_1^* \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \beta \end{bmatrix} - \begin{bmatrix} y_{A_{11}} + y_{P_{11}} & y_{A_{21}} + y_{P_{12}} & 0 \\ y_{A_{12}} + y_{P_{12}} & y_{A_{22}} + y_{P_{22}} & 0 \\ y_{P_{13}} & y_{P_{23}} & 0 \end{bmatrix} \begin{bmatrix} e_1^* \\ e_2^* \\ 0 \end{bmatrix} \tag{2.85}$$

The second line of this equation forces $y_{P_{12}}$ and $y_{P_{22}}$ such that

$$(y_{A_{12}} + y_{P_{12}})\, e_1^* + (y_{A_{22}} + y_{P_{22}})\, e_2^* = 0 \tag{2.86}$$

Once solved, the optimal source impedance is

$$y_{S_{Opt}} = -(y_{A_{11}} + y_{P_{11}}) - (y_{A_{21}} + y_{P_{12}}) \frac{e_2^*}{e_1^*} \tag{2.87}$$

Note that for each value of $y_{P_{13}}$ and $y_{P_{23}}$, there is a β that satisfies the minimum noise measure. Therefore, these two y-parameters can be optimized such that the calculated optimal source impedance also provides the correct input impedance for power matching. Unfortunately, the required $y_{P_{12}}$ and $y_{P_{22}}$ may result in an unstable amplifier, limiting the use of this technique.

2.6 Optimal Bias Condition

Considering that the minimum noise figure and the maximum available power gain together play a crucial role in the performance of an LNA, neither of them should be considered alone

to find the optimal bias condition. Moreover, the optimal source impedance for the minimum noise figure and maximum available gain may be different due to the noise correlation at different ports. For this reason, any noise measure proxy such as

$$M_{proxy} = \frac{NF_{min} - 1}{1 - \frac{1}{G_{max}}} \tag{2.88}$$

will not be correct. Fig. 2.11 shows that the minimum noise measure is obtained at about

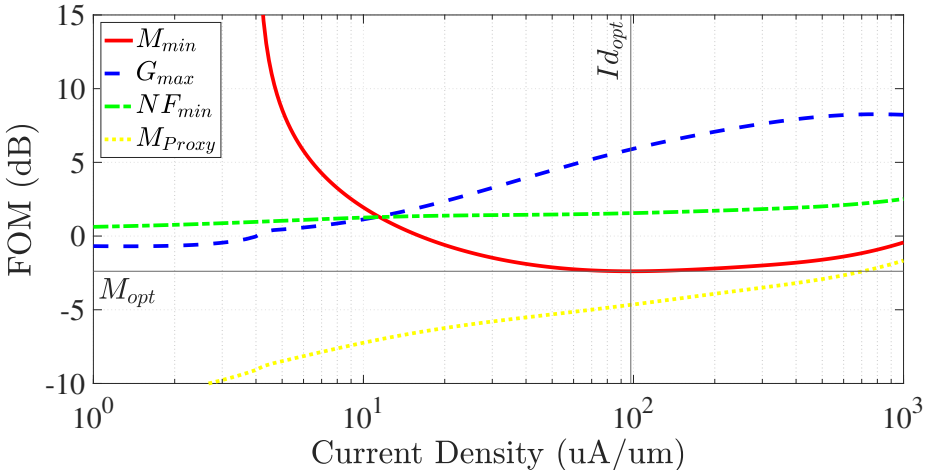


Figure 2.11: Different FOMs vs. bias current density

100μA/μm current density. Note that in some cases, the BSIM noise models show a wrong trend [27], which is not physically possible. Therefore, the optimal current density should be chosen based on the measured data.

Note that the optimal current density calculated here is lower than most typical LNAs. This is because it assumes that a loss-less matching network can be implemented. However, in typical millimeter-wave amplifiers, matching networks contribute 1 to 3dB to the insertion loss. Note that the noise measure of cascaded amplifiers can be derived as

$$M_C = M_1 + (M_2 - M_1)\frac{G_2 - 1}{G_1 G_2 - 1} \tag{2.89}$$

By definition, the noise measure of a passive lossy device is equal to -1. Therefore, the noise measure of the amplifier with its matching network can be calculated as

$$M_C = M_1 + (M_1 + 1)\frac{1 - IL}{G_1 IL - 1} \tag{2.90}$$

where IL is the insertion loss of the matching network. Note that as the gain of the amplifier decreases, even if the noise measure for the active device is the same, the noise measure of the cascade decreases. Therefore, the optimal current density should be chosen in an iterative process. For most of the low-noise amplifiers implemented in this work, a current density of $\approx 200\mu$A/μm is used.

2.A Simulation Flow of Minimum Noise Measure

Since most simulation tools do not directly calculate the noise measure, this method is used with Spectre.

1. The bias circuit of the device under test is present.
2. Ports with 50Ω internal impedance are added.
3. Noise generation of all ports is enabled, and the noise temperature is explicitly set to the simulation temperature.
4. The S-parameter simulation is set to the correct frequency range.
5. Two sets of output files are generated:
   - Y-parameters: With the data format set to "touchstone", the parameter type set to "y", and the noise data set to "no", the normalized y-parameters of the device under test are calculated and extracted. This file can be read immediately by the CAD tools.
   - Noise Cross-Correlation: with the data format set to "Spectre", the parameter type set to "y", and the noise data set to "cy", the normalized noise cross-correlation matrix of the device under test is calculated and extracted. Since the noise cross-correlation matrix is a Hermitian matrix, only half of the entries are exported: diagonal values with a single real number and off-diagonal values with a pair of real and imaginary numbers. Since this format is not necessarily compatible with CAD tools, the extracted values must be put into a suitable format (e.g., CSV) and then imported.
6. Y-parameter and noise cross-correlation files are imported. Since each of these files is normalized, the normalization factors should be considered.
   - The Y-parameters are normalized to $(50\Omega)^{-1}$, and therefore all entries of the Y-matrix should be multiplied by $(50\Omega)^{-1}$.
   - The noise cross-correlation matrix is normalized to $4\mathrm{kT}\Delta f$, where T is the port temperature and not the simulation temperature. Note that the characteristic noise matrix must also be normalized by a factor of $2\mathrm{kT}\Delta f$.
7. The following equation can be used to derive the characteristic matrix:

$$\mathbf{N} = -2 \times \left(\mathbf{Y}_A + \mathbf{Y}_A{}^H\right)^{-1} \left(\frac{\overline{\mathbf{I}_N \mathbf{I}_N{}^H}}{4\mathrm{kT}\Delta f}\right) \tag{2.91}$$

8. The eigenvalues of the characteristic noise matrix are calculated, and the smallest positive value is taken as the minimum noise measure. If there is no positive eigenvalue, it can be concluded that the power gain is less than 0dB.

Chapter 3

140GHz Receiver Design

In this chapter, a wideband receiver at 140GHz is explained. Note that the carrier frequency is close to the $\frac{f_t}{2}$ [1], so the receiver chain should be carefully optimized to get the most out of the available technology.

Fig. 3.1 shows the block diagram of the receiver. Each section is carefully examined in the remainder of this chapter.

3.1 Low-Loss LC Matching Networks

Before implementing the receiver, it is instructive to study the behavior of the matching network since the insertion loss of the matching network is not negligible at millimeter-wave frequencies. The insertion loss of the matching network in Fig. 3.2 is

$$IL = \frac{P_L}{P_L + P_M} \tag{3.1}$$

Under the assumption of series matching [2]

$$\begin{aligned} IL &= \frac{I^2 R_L}{I^2 (R_L + R_M)} \\ &= \frac{R_L}{R_L + R_M} \end{aligned} \tag{3.2}$$

[1] f_t is the unity current-gain frequency.

[2] Without loss of generality, parallel elements exhibit the same behavior.

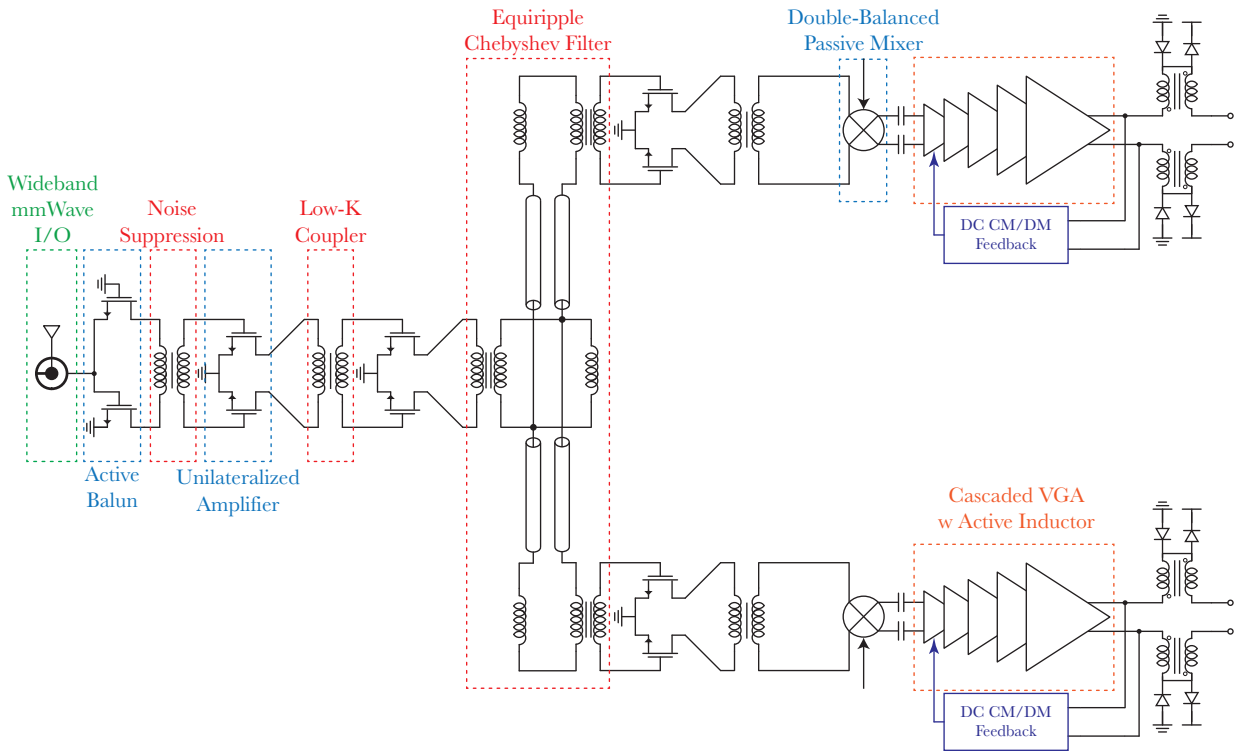


Figure 3.1: Block diagram of the receiver chain

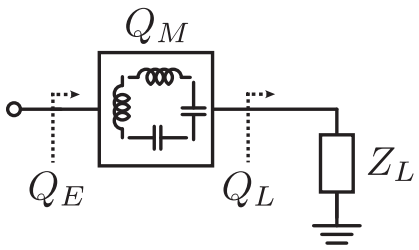


Figure 3.2: Single-component matching network

where R_L, X_L, R_M, and X_M are resistance and reactance of the load and matching component impedances. Therefore,

$$IL = \frac{R_L}{R_L + \frac{X_M}{Q_M}} \tag{3.3}$$

$$= \frac{R_L}{R_L + \frac{(X_L+X_M)-X_L}{Q_M}} \tag{3.4}$$

$$= \frac{1}{1 + \frac{\frac{X_L+X_M}{R_L} - \frac{X_L}{R_L}}{Q_M}} \tag{3.5}$$

$$= \frac{1}{1 + \frac{\frac{X_L+X_M}{R_L} - Q_L}{Q_M}} \tag{3.6}$$

If the insertion loss were small, it could be easily simplified at this point. However, most on-chip networks have a high loss. The above equation can be written as follows to get the exact formula

$$\begin{aligned} IL &= \frac{1}{1+\frac{\frac{X_L+X_M}{R_L+R_M}\frac{R_L+R_M}{R_L}-Q_L}{Q_M}} \\ &= \frac{1}{1+\frac{Q_E\frac{R_L+R_M}{R_L}-Q_L}{Q_M}} \\ &= \frac{1}{1+\frac{Q_E\frac{1}{IL}-Q_L}{Q_M}} \end{aligned} \tag{3.7}$$

where Q_E is the equivalent quality factor of the impedance seen at the end of the matching network (Fig. 3.2). The exact insertion loss can be derived by solving the previous equation as

$$IL = \frac{Q_M - Q_E}{Q_M - Q_L} \tag{3.8}$$

$$= \frac{1-\frac{Q_E}{Q_M}}{1-\frac{Q_L}{Q_M}} \tag{3.9}$$

Note that in a low-loss network, where $Q_M \gg Q_E$ and $Q_M \gg Q_L$, the insertion loss can be calculated approximately as

$$IL = 1 - \frac{Q_E - Q_L}{Q_M} \tag{3.10}$$

$$= \frac{1}{1+\frac{Q_E-Q_L}{Q_M}} \tag{3.11}$$

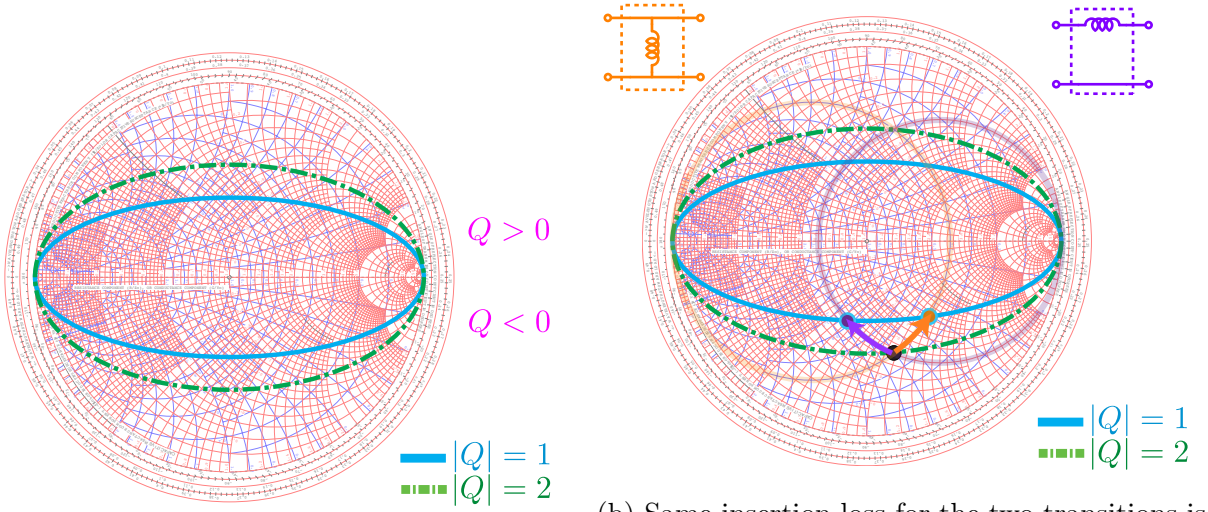
Note that in the above equations, the quality factors are defined as $Q_M = \frac{X_M}{R_M}$, which means that the quality factor of a capacitor is negative; and the quality factor of an inductor is positive, as in Fig. 3.3a. From this, we conclude that

$$Q_L < Q_M \Rightarrow Q_E < Q_M \tag{3.12}$$

$$Q_M < Q_L \Rightarrow Q_M < Q_E \tag{3.13}$$

When using multiple matching components as in Fig. 3.4,

$$\begin{aligned} IL =& \frac{P_L}{P_L + P_{M1} + P_{M2}} \\ =& \frac{P_L + P_{M1}}{P_L + P_{M1} + P_{M2}} \frac{P_L}{P_L + P_{M1}} \\ =& IL_2 \times IL_1 \\ =& \frac{Q_{M2} - Q_{E2}}{Q_{M2} - Q_{E1}} \frac{Q_{M1} - Q_{E1}}{Q_{M1} - Q_L} \end{aligned} \tag{3.14}$$



(a) The contours of the constant Q

(b) Same insertion loss for the two transitions is expected

Figure 3.3: Definition of Q and moving between different Q-contours.

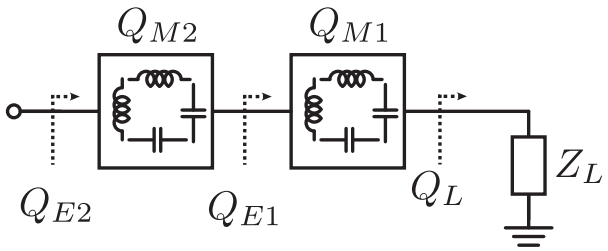


Figure 3.4: Cascade of several elements

if $Q_{M1} = Q_{M2} = Q_M$

$$IL = \frac{Q_M - Q_{E2}}{Q_M - Q_L} \tag{3.15}$$

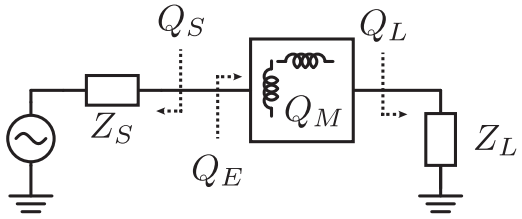


Figure 3.5: Circuit model used to obtain the maximum transmission

Let us now design a matching network for an amplifier with circuit model shown in Fig. 3.5. Let us assume that it is a unilateral amplifier

$$G_{tot} = G_{mux} \times \left(1 - |\Gamma|^2\right) \times IL \tag{3.16}$$

$$T = \left(1 - |\Gamma|^2\right) \times IL \tag{3.17}$$

A suitable matching network should maximize T. In a loss-less system ($IL = 1 = 0$dB), maximum transmission is achieved when reflection is minimized. However, if we consider a lossy network, the optimum looks different. Note that $1-|\Gamma|^2$ represents the accepted power normalized to the available power

$$1 - |\Gamma|^2 = \frac{P_S}{P_{S,max}} \tag{3.18}$$

$$P_{S,max} = \frac{{V_S}^2}{(2R_S)^2} R_S \tag{3.19}$$

$$P_S = \frac{{V_S}^2}{\left|R_S(1+jQ_S) + R_E(1+jQ_E)\right|^2} R_E \tag{3.20}$$

$$1 - |\Gamma|^2 = \frac{\frac{{V_S}^2}{\left|R_S(1+jQ_S)+R_E(1+jQ_E)\right|^2} R_E}{\frac{{V_S}^2}{(2R_S)^2} R_S} \tag{3.21}$$

$$= \frac{4R_S R_E}{\left|R_S(1+jQ_S) + R_E(1+jQ_E)\right|^2} \tag{3.22}$$

$$= \frac{4R_S R_E}{\left(R_S + R_E\right)^2 + \left(R_S Q_S + R_E Q_E\right)^2} \tag{3.23}$$

$$T = \frac{4R_S R_E}{\left(R_S + R_E\right)^2 + \left(R_S Q_S + R_E Q_E\right)^2} \frac{Q_M - Q_E}{Q_M - Q_L} \tag{3.24}$$

Let us first consider the case where the loss of the matching network is negligible. In this case, the conjugate matching condition yields maximum transmission when $Q_E = -Q_S$ and $R_S = R_E$. The assumption of a low-loss matching network holds as long as

$$\left|\frac{Q_S}{Q_M}\right| \ll 1, \left|\frac{Q_L}{Q_M}\right| \ll 1 \Rightarrow IL \approx \left(1 + \frac{Q_S}{Q_M}\right)\left(1 + \frac{Q_L}{Q_M}\right) \tag{3.25}$$

Suppose that the quality factor of the source or load is comparable to the magnitude of the quality factor of the components of the matching network. In this case, the conjugate matching does not provide the maximum transmission. To achieve the maximum transmission

$$\frac{\partial T}{\partial R_E} = 0 \Rightarrow R_E = R_S \sqrt{\frac{1 + {Q_S}^2}{1 + {Q_E}^2}} \tag{3.26}$$

$$\frac{\partial T}{\partial Q_E}\bigg|_{R_E=R_S\sqrt{\frac{1+{Q_S}^2}{1+{Q_E}^2}}} = 0 \Rightarrow Q_{E,opt} = -Q_S + \frac{2Q_M({Q_S}^2+1)}{1+2Q_M Q_S - {Q_M}^2} \tag{3.27}$$

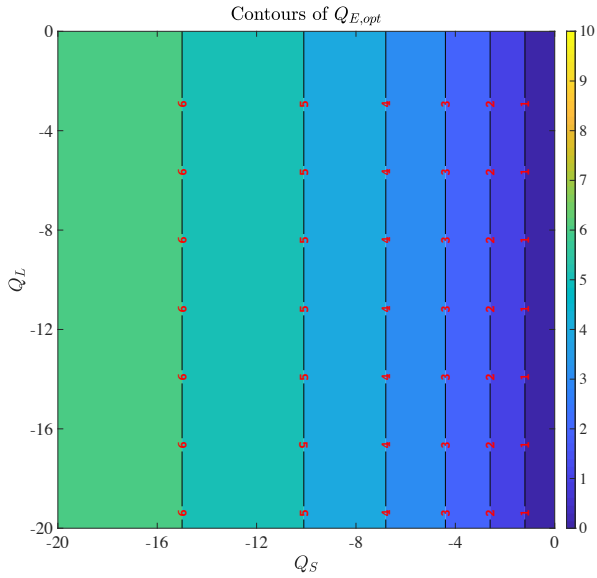


Figure 3.6: The optimal input quality factor for the network with $Q_M = 20$.

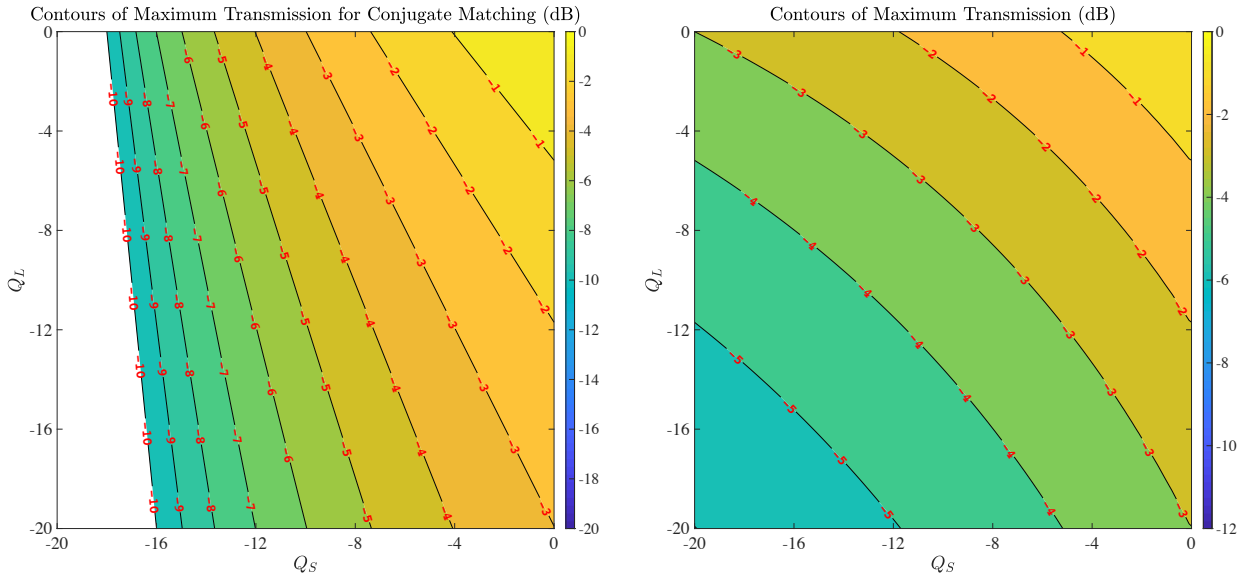
As shown in Fig. 3.6, the optimal quality factor is different from the conjugate matching condition. Assuming a reasonable passive component $|Q_M| > 1$

$$T_{opt} = \frac{{Q_M}^2+1}{(Q_M - Q_S)(Q_M - Q_L)} \tag{3.28}$$

$$R_{E,opt} = R_S - \frac{2(Q_M Q_S + 1)}{{Q_M}^2+1} R_S \tag{3.29}$$

$$X_{E,opt} = -X_S - \frac{2(Q_M - Q_S)}{{Q_M}^2+1} R_S \tag{3.30}$$

Fig. 3.7a shows the transmission loss in a lossy matching network. Note that the insertion loss for source and load quality factors is not symmetric. This may seem unreasonable and counterintuitive. To understand this problem, consider Fig. 3.8, in which $Z_L = 1 - j5\Omega$ is matched to $Z_S = 2 - j5\Omega$ using a lossy inductor with $Z_M = 1 + j10\Omega$. In this simple schematic, the input impedance on the source side is $Z_{E,Source} = 1 + j10 + 1 - j5 = 2 + j5\Omega$, which provides a perfect conjugate match on the source side and eliminates any reflections ($\Gamma_S = 0$). On the other hand, the output impedance on the load side is $Z_{E,Load} = 1 + j10 + 2 - j5 = 3 + j5\Omega$. Although the termination on the source side is matched, the load impedance sees an unmatched termination with a reflection of $\Gamma_L = \frac{1}{2} = -6\text{dB}$. Therefore, the asymmetry of Fig. 3.7a is due to the choice of which port is matched and which port has nonzero reflection.



(a) Conjugate matching condition (b) Optimal design with reflection

Figure 3.7: Contours of total transmission loss for different source and load quality factors with $Q_M = 20$

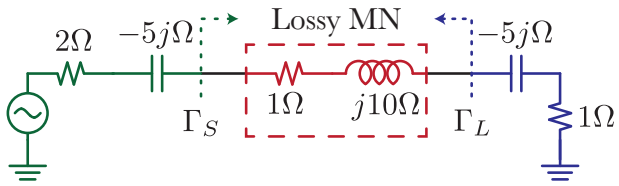


Figure 3.8: Asymmetry of source and load reflections of a lossy matching network

3.2 Transformers

Transformers are popular at millimeter-wave frequencies. Let us study their performance and compare them with LC ladder networks. For a lossy transformer,

$$Z = \begin{bmatrix} R_p + j\omega L_p & j\omega M \\ j\omega M & R_s + j\omega L_s \end{bmatrix} \tag{3.31}$$

Parameter	Conjugate Matched	Transmission Optimized
Resistance	R_S	$R_S - \frac{2(Q_M Q_S+1)}{{Q_M}^2+1} R_S$
Reactance	$-X_S$	$-X_S - \frac{2(Q_M-Q_S)}{{Q_M}^2+1} R_S$
Transmission Loss	$\frac{Q_M+Q_S}{Q_M-Q_L}$	$\frac{{Q_M}^2+1}{(Q_M-Q_S)(Q_M-Q_L)}$
Optimum Q_E	$-Q_S$	$-Q_S + \frac{2Q_M({Q_S}^2+1)}{1+2Q_M Q_S-{Q_M}^2}$
Comments	Impractical when $\|Q_M\| < \|Q_S\|$	-

Table 3.1: Summery of different matching network design methodologies

Assuming $Z_{jk} = m_{jk} + in_{jk}$, the stability K-factor can be calculated as

$$K = \frac{2m_{11}m_{22} - P}{L} \tag{3.32}$$

$$= \frac{2R_pR_s + \omega^2M^2}{\omega^2M^2} \tag{3.33}$$

$$= 1 + \frac{2R_pR_s}{\omega^2M^2} \tag{3.34}$$

$$= 1 + \frac{2}{k^2Q_pQ_s} \tag{3.35}$$

where $Z_{12}Z_{21} = P + jB = |L|e^{j\theta}$ and $M = k\sqrt{L_pL_s}$, $Q_p = \frac{\omega L_p}{R_p}$, $Q_s = \frac{\omega L_s}{R_s}$

$$G_{max} = \frac{1}{K + \sqrt{K^2 - 1}} = K - \sqrt{K^2 - 1} \tag{3.36}$$

$$= 1 + \frac{2}{k^2 Q_p Q_s} - 2\sqrt{\frac{1}{k^2 Q_p Q_s}\left(1 + \frac{1}{k^2 Q_p Q_s}\right)} \tag{3.37}$$

$$= \frac{1}{1 + \frac{2}{k^2 Q_p Q_s} + 2\sqrt{\frac{1}{k^2 Q_p Q_s}\left(1 + \frac{1}{k^2 Q_p Q_s}\right)}} \tag{3.38}$$

$$= \frac{k^2 Q_p Q_s}{k^2 Q_p Q_s + 2 + 2\sqrt{k^2 Q_p Q_s + 1}} \tag{3.39}$$

$$= \frac{\sqrt{k^2 Q_p Q_s + 1} - 1}{\sqrt{k^2 Q_p Q_s + 1} + 1} \tag{3.40}$$

Based on the equations of [28], the optimal terminations on each side of the transformer can be calculated as

$$Z_{1,opt} = m_{11}\Delta + j\left[\frac{B}{2m_{22}} - n_{11}\right] \tag{3.41}$$

where

$$\Delta = \sqrt{1 - \frac{P}{m_{11}m_{22}} - \left(\frac{B}{2m_{11}m_{22}}\right)^2} \tag{3.42}$$

Using the above equations, the optimal impedance can be calculated as

$$Z_{p,opt} = \sqrt{k^2 Q_p Q_s + 1} R_p - j\omega L_p \tag{3.43}$$

$$= \frac{\sqrt{k^2 Q_p Q_s + 1}}{Q_p}\omega L_p - j\omega L_p \tag{3.44}$$

Similarly,

$$Z_{s,opt} = \sqrt{k^2 Q_p Q_s + 1} R_s - j\omega L_s \tag{3.45}$$

$$= \frac{\sqrt{k^2 Q_p Q_s + 1}}{Q_s}\omega L_s - j\omega L_s \tag{3.46}$$

These optimal impedances are shown in Fig. 3.5. It is instructive to see whether or not the transformer outperforms the LC ladder networks in terms of transmission losses. Note that for a conjugate matching condition, the source and load quality factors (Q_S and Q_L) can be calculated based on the coupling factor (k) and the primary and secondary quality factors

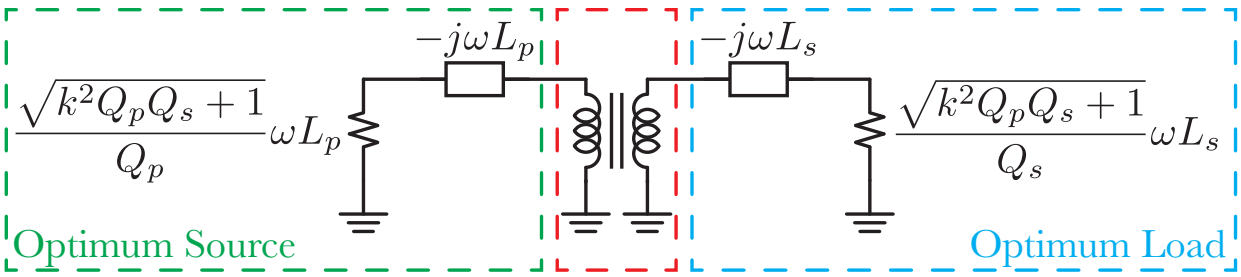


Figure 3.9: Optimal loading condition to achieve the minimum insertion loss of the transformer

(Q_p and Q_s) as

$$Q_S = -\frac{Q_p}{\sqrt{k^2Q_pQ_s+1}} \tag{3.47}$$

$$Q_L = -\frac{Q_s}{\sqrt{k^2Q_pQ_s+1}} \tag{3.48}$$

If LC ladder networks were used,

$$IL_{LC} = \frac{Q_M + Q_S}{Q_M - Q_L} \tag{3.49}$$

$$= \frac{Q - \frac{Q}{\sqrt{k^2Q^2+1}}}{Q + \frac{Q}{\sqrt{k^2Q^2+1}}} \tag{3.50}$$

$$= \frac{\sqrt{k^2Q^2+1}-1}{\sqrt{k^2Q^2+1}+1} \tag{3.51}$$

where the same quality factors $Q_M = Q_p = Q_s = Q$ are considered for all inductors for a fair comparison. Note that the insertion loss of an LC ladder network is the same as that of its transformer counterpart. There are mainly two factors that determine which matching strategy is better. First, in a conjugate matched circuit with an LC ladder network, comparable source and load quality factors are not required for the minimum insertion loss. On the other hand, if impedance transformation is the goal, an optimally matched transformer provides an impedance transformation of

$$\left.\frac{R_S}{R_L}\right|_{Transformer} = \frac{L_s}{L_p} \tag{3.52}$$

However, in an LC ladder network without additional capacitors, there is a minimum and a maximum impedance that can be achieved with step-up or step-down networks (Fig. 3.10) which is

$$\frac{1}{1+{Q_S}^2} < \left.\frac{R_S}{R_L}\right|_{LC} < 1+{Q_L}^2 \tag{3.53}$$

Beyond this range, additional capacitors are required in the matching network, and the additional inductive energy resonating with the new capacitive energy increases the power dissipation. Therefore, when the source and load have a relatively low quality factor but

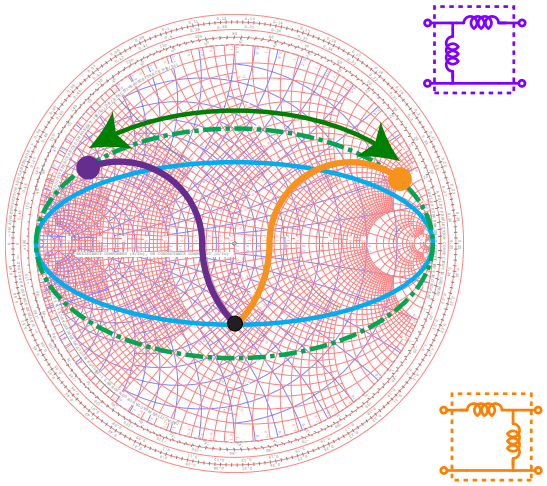


Figure 3.10: Step-up and step-down matching networks with the same insertion loss

a high impedance transformation is required, transformers are superior to their LC ladder counterparts.

In a CMOS process, neglecting gate-drain capacitance,

$$Q_{Gate} = \frac{-1}{R_g \omega C_g} \tag{3.54}$$

$$Q^*_{Drain} = R_d \omega C_d \tag{3.55}$$

where R_g, C_g, R_d and C_d are the gate series resistance, gate capacitance, drain output resistance and drain capacitance, respectively. As the frequency increases, the gate quality factor increases while the gate quality factor decreases (Fig. 3.11). While they are completely different at RF frequencies, these two quality factors become comparable in the millimeter-wave range. Therefore, transformers can be used for matching between stages.

Note that assuming high quality factor transformers $k^2 Q_p Q_s \gg 1$, the optimal termination quality factor approaches $\left|\frac{1}{K}\right|$ when assuming similar quality factors for primary and secondary coils. Therefore, for a conjugate matched network,

$$-Q_S = \frac{1}{|k|} \tag{3.56}$$

For example, if the drain and gate quality factors are $Q_S = Q_g = Q_d = -10$, a transformer with a coupling factor of $|k| = 0.1$ is required. However, as mentioned earlier, the optimal

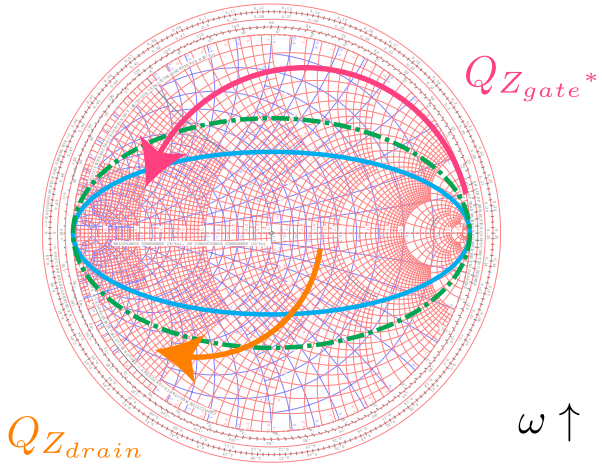
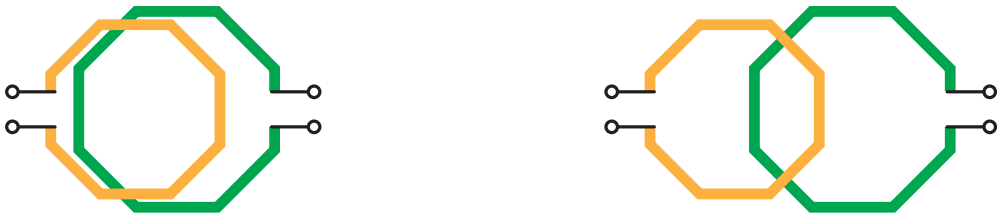


Figure 3.11: Gate and drain quality factor vs. frequency

transmission does not occur under conjugate matched conditions. Using the equations for optimal transmission

$$-Q_S + \frac{2Q_M({Q_S}^2+1)}{1+2Q_M Q_S - {Q_M}^2} = \frac{1}{|k|} \tag{3.57}$$

If transformers with quality factors of $Q_M = Q_p = Q_s = 30$ were used in the previous example, a transformer with a coupling factor of $|k| = 0.16$ provides optimal transmission. Given the low coupling factor of the transformer, its physical shape can be optimized to achieve the highest quality factor possible with the technology. Fig. 3.12a shows how high-k transformers are typically implemented. Note that two thick metal layers are required if no bridges are used. The coupling factor can be reduced by moving the two loops away from each other (Fig. 3.12b).

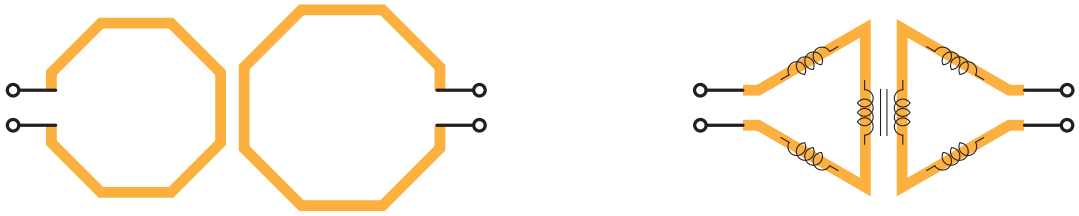


(a) High-k (b) Low-k

Figure 3.12: Symbolic structure of a stacked single-turn transformer

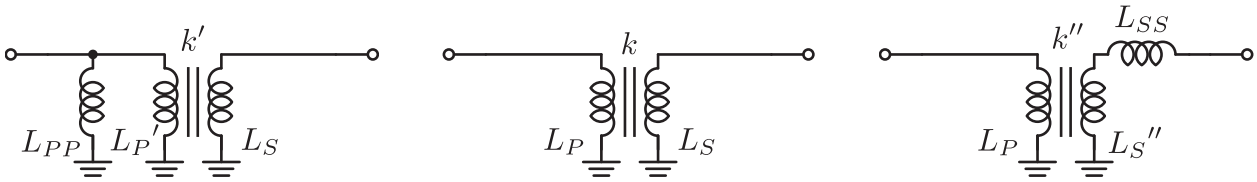
If a low coupling is desired, the transformer could be implemented with a single thick metal layer, as in Fig. 3.13a, where two single inductors are broadside-coupled. With octagonal loops, a maximum coupling of $|k| = \frac{1}{8} = 0.12$ can be achieved since only one of eight

edges is coupled. When triangular loops are used, as in Fig. 3.13b, coupling factors as high as $k = 0.3$ can be achieved when $L_s = L_p$.



(a) Low-k (b) High-k

Figure 3.13: The symbolic structure of a transformer with broadside coupling



(a) Parallel equivalent (b) Reference (c) Series equivalent

Figure 3.14: Transformer equivalents

In some situations, primary and secondary coils must differ for impedance transformation while maintaining moderate to high coupling factors. At RF frequencies, this can be easily accomplished by using an inductor with multiple turns stacked over a single-turn inductor. At millimeter-wave frequencies, the self-resonance-frequency of the transformer prohibits the use of multi-turn inductors. In this case, transformer equivalents can be used, as in Fig. 3.14. In the series equivalent of Fig. 3.14c, where

$$L_S'' + L_{SS} = L_S \tag{3.58}$$

the new transformer has the same Z-matrix as the reference transformer if

$$\frac{M}{(L_P - M) + M} = \frac{M''}{(L_P - M'') + M''} \tag{3.59}$$

$$k\sqrt{L_P L_S} = k''\sqrt{L_P L_S''} \tag{3.60}$$

$$k'' = k\sqrt{\frac{L_S}{L_S''}} = k\sqrt{\frac{L_S}{L_S - L_{SS}}} \tag{3.61}$$

Similarly, in the parallel equivalent of Fig. 3.14a, where

$$L_P' || L_{PP} = L_P \tag{3.62}$$

the new transformer has the same Z-matrix as the reference transformer if

$$\frac{M}{(L_P - M) + M} = \frac{M'}{(L'_P - M') + M'} \tag{3.63}$$

$$\frac{k\sqrt{L_P L_S}}{L_P} = \frac{k'\sqrt{L'_P L_S}}{L'_P} \tag{3.64}$$

$$k' = k\sqrt{\frac{L'_P}{L_P}} = k\sqrt{\frac{L_{PP}}{L_{PP} - L_P}} \tag{3.65}$$

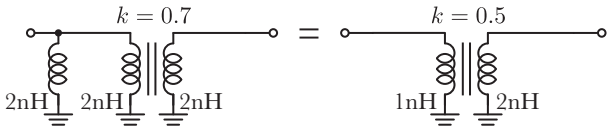
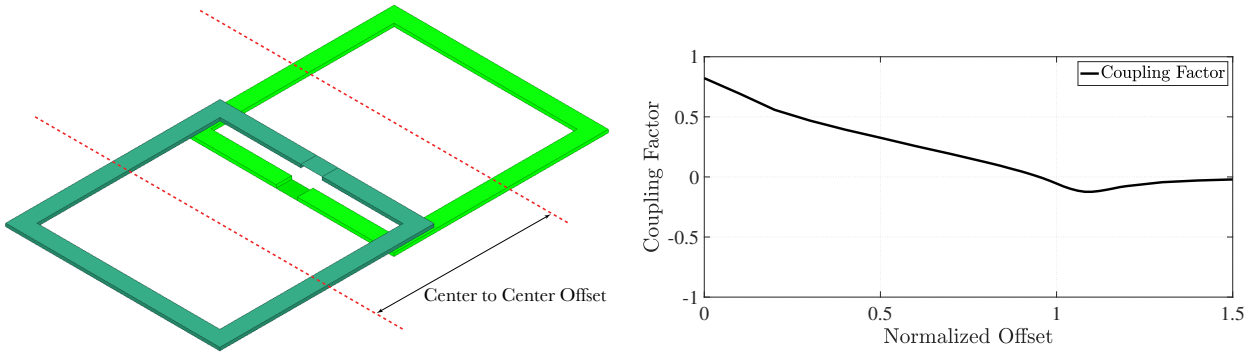


Figure 3.15: Example of equivalent transformer topologies

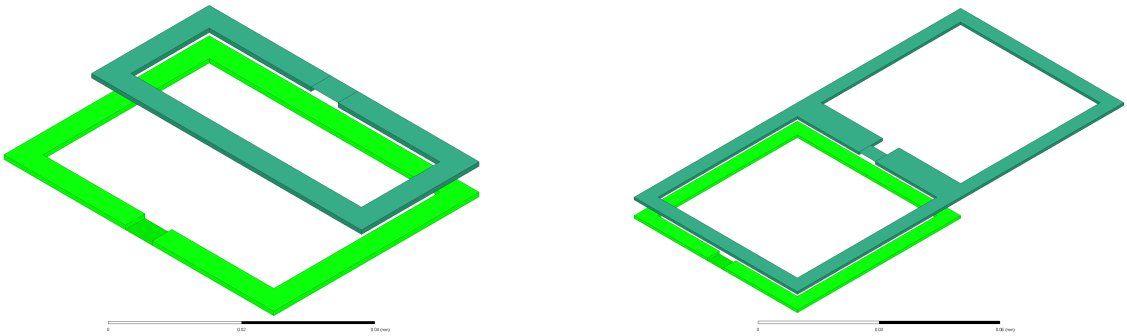
To show the effectiveness of this method, consider the example in Fig. 3.15. It is difficult to achieve a coupling factor of 0.5 when the secondary inductance is twice the primary inductance. As suggested by [29], the inductance of spiral inductors is directly related to the length of the inductor, typically around 1pH/um as a rule of thumb.



(a) Spiral transformer (b) Coupling factor

Figure 3.16: Coupling factor of a transformer as a function of the center-to-center distance, normalized to the width of each loop

On the other hand, the coupling factor is usually determined by the mutual inductance of the parallel legs of each inductor. As you can see in Fig. 3.16b, the mutual inductance decreases as the two inductors move apart because the length of the overlapping side metals decreases. Note that the coupling through the lateral runs opposes the end legs. Fig. 3.16a



(a) Single Loop (b) Dual Loop

Parameter	Single Loop	Dual Loop
Primary Inductance	103pH	104pH
Secondary Inductance	69pH	69pH
Coupling Factor	.34	.51

Figure 3.17: Increasing the coupling factor by using transformer equivalents

shows the exact offset at which they cancel each other. Beyond this point, the lateral coupling is negligible, and the coupled current flows in the opposite direction. Intuitively, higher mutual inductance can be achieved by using larger loops for the primary or secondary. However, this is accompanied by higher inductance for the corresponding loop, which can be compensated by adding series or parallel inductance. Fig. 3.17 shows simulation results demonstrating the effectiveness of this technique to increase the effective coupling factor.

3.3 High Quality-Factor Inductors

As described in the previous section, high-quality inductors are required to minimize the insertion loss of the amplifier. Most inductors are designed as a single-turn loop in the millimeter-wave range to achieve a high self-resonance frequency (SRF). Assuming that a single-turn inductor can be modeled as a lossy transmission line, its impedance can be described as

$$Z_L = Z_0 \frac{1 - e^{-2\gamma d}}{1 + e^{-2\gamma d}} \tag{3.66}$$

where γ is the propagation constant and d is the length of the transmission line. The propagation constant can be written as

$$\gamma = \sqrt{(i\omega L' + R')(i\omega C' + G')} \tag{3.67}$$

$$\approx i\omega\sqrt{L'C'}\left(1 + \frac{R'}{2i\omega L'} + \frac{G'}{2i\omega C'}\right) \tag{3.68}$$

where L', R', C' and G' are respectively the inductance, series resistance, capacitance and shunt conductance per unit length. Note that the second approximation applies only to low-loss structures. The quality factor of the inductance is

$$Q_L = \frac{\mathfrak{Im}\{Z_L\}}{\mathfrak{Re}\{Z_L\}} \tag{3.69}$$

$$\approx \frac{2e^{-\sqrt{L'C'}d\left(\frac{R'}{L'}+\frac{G'}{C'}\right)}}{1-e^{-2\sqrt{L'C'}d\left(\frac{R'}{L'}+\frac{G'}{C'}\right)}}\sin(2d\omega\sqrt{L'C'}) \tag{3.70}$$

Note that the maximum inductance is reached when $2d\omega\sqrt{L'C'} = \frac{\pi}{2}$ and the peak quality factor is

$$Q_{L,max} \approx \frac{1}{\sqrt{L'C'}d\left(\frac{R'}{L'}+\frac{G'}{C'}\right)} \tag{3.71}$$

Note that the peak quality factor is only a function of the length of the inductor. The inductance of the loop can be calculated as follows

$$\mathfrak{Im}\{Z_L\} \approx Z_0\frac{2e^{-\sqrt{L'C'}d\left(\frac{R'}{L'}+\frac{G'}{C'}\right)}\sin(2d\omega\sqrt{L'C'})}{1+2e^{-\sqrt{L'C'}d\left(\frac{R'}{L'}+\frac{G'}{C'}\right)}\cos(2d\omega\sqrt{L'C'})+e^{-2\sqrt{L'C'}d\left(\frac{R'}{L'}+\frac{G'}{C'}\right)}} \tag{3.72}$$

For the peak quality factor, the inductance of the loop seems to be independent of the frequency and equal to

$$\mathfrak{Im}\{Z_L\}|_{2d\omega\sqrt{L'C'}=\frac{\pi}{2}} \approx Z_0 \tag{3.73}$$

The definition of characteristic impedance is not clear here. Note that as the loop diameter increases, the characteristic impedance also increases. The consequence of this trend is that a higher optimum inductance can be expected in a lower frequency range. However, as the frequency decreases, the conductivity of the substrate (σ) dominates over its permittivity, as

$$\epsilon_c(\omega) = \epsilon_r\epsilon_0 - i\frac{\sigma}{\omega} \tag{3.74}$$

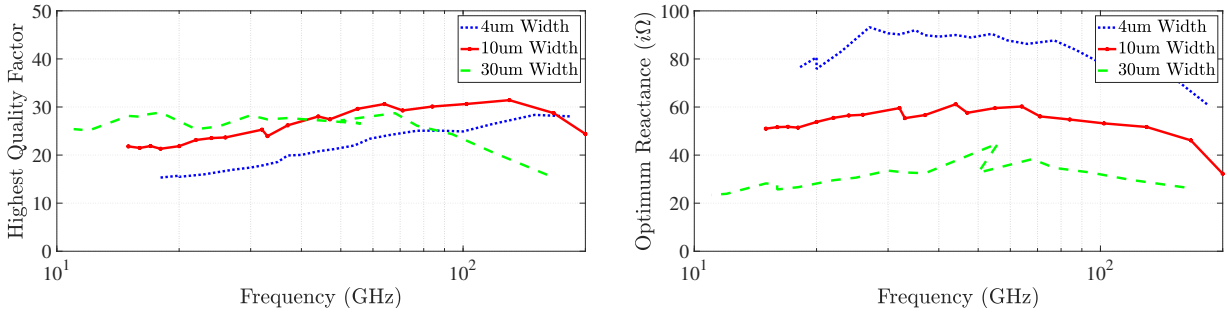
and therefore, the transmission line model in this section resembles a differential microstrip line. For a low-doped silicon with a conductivity of $10\Omega^{-1}\mathrm{m}^{-1}$, this transition occurs around 15GHz. Since the operating frequency of this work is much higher than 15GHz, a quasi-TEM wave is considered for the twinstrip line ([30]). The characteristic impedance of a homogeneous twinstrip line can be approximately calculated as follows

$$Z_{Twin} \approx \sqrt{\frac{\mu}{\epsilon\epsilon_r}}\frac{1}{\pi}\cosh^{-1}\left(1+\frac{S}{W}\right) \tag{3.75}$$

Here S is the distance between the strips and W is the width of each strip. Note that the above equation can be approximated as follows when the distance between the strips is much larger than the width of each strip

$$Z_{Twin} \approx \sqrt{\frac{\mu}{\epsilon\epsilon_r}}\frac{1}{\pi}\ln\left(1+\frac{S}{W}\right) \tag{3.76}$$

This means that the characteristic impedance becomes a weak function of the spacing. Calculating the relative dielectric constant requires conformal mapping, which is beyond the scope of this chapter. Instead, an average dielectric constant of the silicon and interlayer dielectric can be considered. Simulation results show that the optimal reactance depends to some extent on the width of the inductor but is relatively independent of the loop diameter. However, decreasing the width of the inductor may increase the optimal impedance at the expense of a lower quality factor. Fig. 3.18 summarizes the simulation results for different inductor widths.



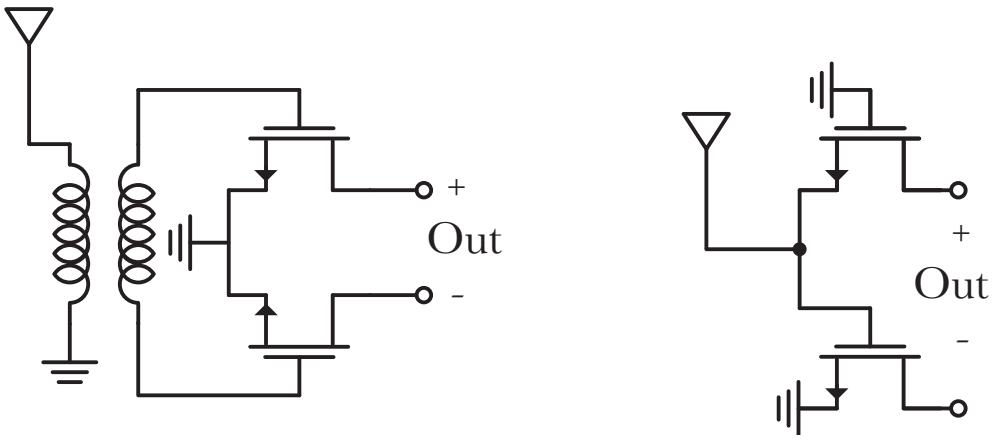
(a) Inductor reactance (b) Quality factor

Figure 3.18: Optimal inductors at different frequencies

The conclusion is that the designer should know the range of optimum reactances when high-quality inductors are required. Since transformers consist of coupled inductors, the same argument applies to them. Transformer equivalents should be used if primary and secondary inductors deviate from the optimum reactances.

3.4 Low Noise Active Balun

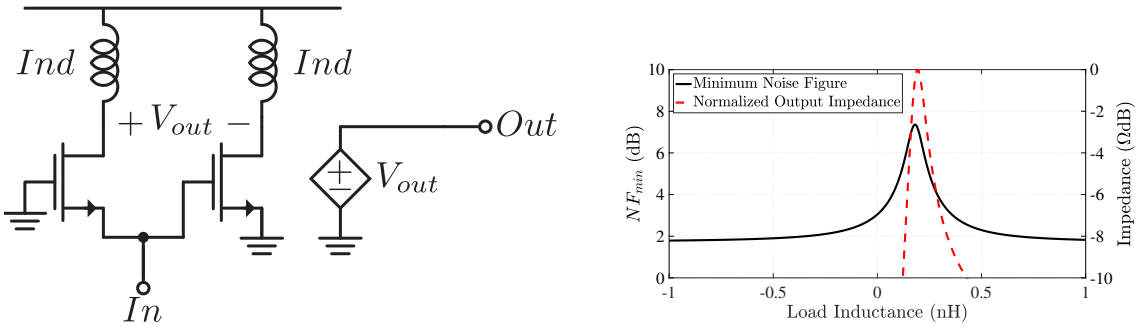
Conventionally, passive baluns (Fig. 3.19a) are used to convert single-ended signals coming from the antenna into differential signals before passing them to low-noise differential amplifiers. These passive baluns are lossy and contribute to a noise figure of about 2dB. As an alternative, single-ended LNAs can be used that do not require conversion of single-ended to differential signals, saving about 2dB of noise degradation. However, electromagnetic modeling of single-ended amplifiers is complicated, and designers tend to worry about the



(a) Passive (b) Active

Figure 3.19: Active and passive balun topologies

possibility of oscillations due to unpredictable instabilities. Therefore, despite the merits of single-ended LNAs, most mmWave LNAs are preceded by a passive balun. As mentioned earlier, the minimum achievable noise measure does not change when more active stages are added. Therefore, it can be assumed that using a common-source stage in parallel with a common-gate stage (as in Fig. 3.19b) will still achieve the minimum noise measure for each stage. First, the minimum noise figure of Fig. 3.20a should be examined.



(a) Topology (DC bias not shown) (b) Performance for different inductive loads

Figure 3.20: Active balun with inductive termination

As you can see in Fig. 3.20b, the minimum achievable noise figure changes for different inductive terminations. More importantly, the minimum noise figure peaks for the inductive loads that resonate with the output capacitance of the active balun. This dilemma can be investigated using noise measure theory. The simulated 3-port Y-parameters for a post-

extraction core are

$$\mathbf{Y_A} = \begin{bmatrix} 0.0144 + 0.0173i & -0.0002 - 0.0027i & -0.0012 - 0.0032i \\ 0.0118 - 0.0043i & 0.0014 + 0.0060i & -0.0000 + 0.0000i \\ -0.0132 - 0.0016i & 0.0000 + 0.0000i & 0.0014 + 0.0060i \end{bmatrix} \tag{3.77}$$

and the correlation matrix is

$$\mathbf{N}_C = \begin{bmatrix} 0.0106 + 0.0000i & -0.0003 + 0.0010i & -0.0094 - 0.0010i \\ -0.0003 - 0.0010i & 0.0097 + 0.0000i & -0.0000 - 0.0000i \\ -0.0094 + 0.0010i & -0.0000 + 0.0000i & 0.0097 + 0.0000i \end{bmatrix} \tag{3.78}$$

The characteristic noise matrix can be calculated as

$$\mathbf{N} = \begin{bmatrix} -0.8017 + 0.0887i & -0.8490 - 0.0971i & 0.8490 + 0.0971i \\ 3.4938 - 0.0987i & -3.3895 - 0.0929i & -3.6183 + 0.0929i \\ 2.7123 + 0.1871i & -4.4671 - 0.0042i & -2.5406 + 0.0041 \end{bmatrix} \tag{3.79}$$

which has three eigenvalues

$$\lambda_{1,2,3} = \{-7.01, -0.29, 0.566\} \tag{3.80}$$

The smallest positive eigenvalue (λ_3) determines the minimum noise measure of this architecture. As expected, the minimum noise measure remains the same as a single transistor. The eigenvectors can be calculated as

$$\mathbf{V}_{\lambda_1} = \begin{bmatrix} 0.00 - 0.00i \\ 0.71 + 0.00i \\ 0.71 + 0.00i \end{bmatrix} \tag{3.81}$$

$$\mathbf{V}_{\lambda_2} = \begin{bmatrix} 0.77 + 0.00i \\ 0.16 + 0.07i \\ 0.61 - 0.07i \end{bmatrix} \tag{3.82}$$

$$\mathbf{V}_{\lambda_3} = \begin{bmatrix} 0.60 + 0.07i \\ -0.18 + 0.07i \\ 0.78 + 0.00i \end{bmatrix} \tag{3.83}$$

Note that to achieve the minimum noise measure

$$y_S \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}^H \begin{bmatrix} 0.60 + 0.07i \\ -0.18 + 0.07i \\ 0.78 + 0.00i \end{bmatrix}^* = \beta \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} - \left(\mathbf{Y}_P + \mathbf{Y}_A\right)^T \begin{bmatrix} 0.60 + 0.07i \\ -0.18 + 0.07i \\ 0.78 + 0.00i \end{bmatrix}^* \tag{3.84}$$

which can be simplified as

$$\begin{bmatrix} y_s(0.60 - 0.07i) \\ 0 \\ 0 \end{bmatrix} = \beta \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} - \begin{bmatrix} -0.0029 + 0.0081i \\ -0.0002 - 0.0028i \\ 0.0002 + 0.0028i \end{bmatrix} - \mathbf{Y}_P \begin{bmatrix} 0.60 - 0.07i \\ -0.18 - 0.07i \\ 0.78 + 0.00i \end{bmatrix} \tag{3.85}$$

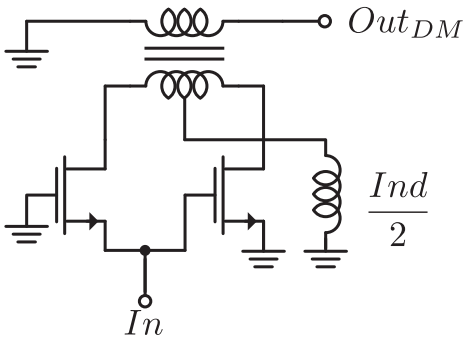
which assumes a reciprocal and symmetric peripheral network that

$$\mathbf{Y}_P = \begin{bmatrix} y_{11} & y_{12} & y_{12} \\ y_{12} & y_{22} & y_{23} \\ y_{12} & y_{23} & y_{22} \end{bmatrix} \tag{3.86}$$

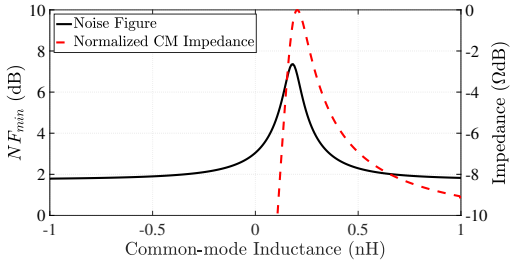
In the absence of a direct path from the input to any of the outputs ($y_{12} = 0$), obtaining the minimum noise measure requires that

$$y_{22} = -y_{23} \tag{3.87}$$

In other words, the passive network should be purely differential at the output. If the symmetry is broken (e.g., by asymmetric passive components or CS and CG stages with different transconductance), the above condition is no longer valid. To prove the hypothesis, we examine the topology of Fig. 3.21a. As you can see in Fig. 3.21b, the common-mode termination indeed changes the differential performance. The critical observation is that the



(a) Topology (DC bias not shown)



(b) Performance for different CM inductive loads

Figure 3.21: Active balun with separation of common-mode and differential-mode terminations

common-mode output impedance of the *peripheral network* itself should be high to achieve the minimum noise measure. As you can see in Fig. 3.21b, the minimum noise figure increases as the inductance values resonate with the common-mode capacitance of the core transistors

Note that the optimum source impedance is achieved when

$$y_s + y_{11} = 0.0064 - 0.0130i \tag{3.88}$$

which has a low quality factor, allowing a low-loss and wideband input matching.

It can be observed that the optimal source is approximately $\frac{1}{2g_m}$. Recall that in the transistor model of Fig. 2.6

$$\begin{bmatrix} i_G \\ i_D \end{bmatrix} = \mathbf{Y}_{CS} \begin{bmatrix} v_G - v_S \\ v_D - v_S \end{bmatrix} \tag{3.89}$$

$$\mathbf{Y}_{CS} = \frac{1}{R_g(C_\pi + C_\mu)S + 1} \begin{bmatrix} (C_\pi + C_\mu)S & -C_\mu S \\ g_m - C_\mu S & (R_g C_\pi S + g_m R_g + 1)\, C_\mu S \end{bmatrix} \tag{3.90}$$

where $\mathbf{Y}_{CS}$ is the Y-parameter of a common-source topology. The Y-parameters of a common-gate topology can be easily computed if we note that

$$\begin{bmatrix} i_S \\ i_D \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} i_G \\ i_D \end{bmatrix} \tag{3.91}$$

and

$$\begin{bmatrix} v_S - v_G \\ v_D - v_G \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} v_G - v_S \\ v_D - v_S \end{bmatrix} \tag{3.92}$$

Therefore,

$$\begin{bmatrix} i_S \\ i_D \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 0 & 1 \end{bmatrix} \mathbf{Y}_{CS} \begin{bmatrix} -1 & 0 \\ -1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} v_S - v_G \\ v_D - v_G \end{bmatrix} \tag{3.93}$$

which means that the Y-parameter of a common-gate topology can be computed as

$$\mathbf{Y}_{CG} = \begin{bmatrix} -1 & -1 \\ 0 & 1 \end{bmatrix} \mathbf{Y}_{CS} \begin{bmatrix} -1 & 0 \\ -1 & 1 \end{bmatrix}^{-1} \tag{3.94}$$

The Y-parameters of the active balun with a CS and a CG stage can be calculated as

$$\mathbf{Y}_{CSCG} = \begin{bmatrix} y_{CG_{11}} + y_{CS_{11}} & y_{CG_{12}} & y_{CS_{12}} \\ y_{CG_{21}} & y_{CG_{22}} & 0 \\ y_{CS_{21}} & 0 & y_{CS_{22}} \end{bmatrix} \tag{3.95}$$

which corresponds to the following formula

$$\begin{bmatrix} i_{in} \\ i_P \\ i_M \end{bmatrix} = \mathbf{Y}_{CSCG} \begin{bmatrix} v_{in} \\ v_P \\ v_M \end{bmatrix} \tag{3.96}$$

Since the optimum noise measure requires the correct choice of common-mode and differential-mode impedance, the above equation can be modified as follows

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} i_{in} \\ i_{DM} \\ i_{CM} \end{bmatrix} = \mathbf{Y}_{CSCG} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 1 \\ 0 & \frac{-1}{2} & 1 \end{bmatrix} \begin{bmatrix} v_{in} \\ v_{DM} \\ v_{CM} \end{bmatrix} \tag{3.97}$$

which means

$$\mathbf{Y}_{CMDM} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & 1 \end{bmatrix}^{-1} \mathbf{Y}_{CSCG} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 1 \\ 0 & \frac{-1}{2} & 1 \end{bmatrix} \tag{3.98}$$

Assuming that the common-mode termination is high, the Y-parameters of the two-port can be calculated as

$$\mathbf{Y}_{DM} = \left(\left(\mathbf{Y}_{CMDM}{}^{-1} \right)_{[1:2,1:2]} \right)^{-1} \tag{3.99}$$

Using a symbolic math package,

$$\begin{aligned} \mathbf{Y}_{DM} = & \frac{1}{R_g(C_\pi + C_\mu)s + 1} \times \\ & \begin{bmatrix} g_m + 2C_\pi s + \dfrac{C_\mu s(C_\pi R_g s + R_g g_m + 1)}{2} & \dfrac{C_\mu s(C_\pi R_g s + R_g g_m - 1)}{2} \\ g_m + \dfrac{C_\mu s(C_\pi R_g s + R_g g_m - 1)}{2} & \dfrac{C_\mu s(C_\pi R_g s + R_g g_m + 1)}{2} \end{bmatrix} \end{aligned} \tag{3.100}$$

Since the output resistance of the devices is neglected, the output impedance at DC approaches infinity. This means that a high passive impedance transformation is required for simultaneous conjugate input-output matching, which increases the noise figure due to the loss of the matching network. However, as the frequency increases, a lower impedance transformation ratio is required. Fig. 3.22 shows the optimum source impedance for the minimum noise figure and maximum available power gain.

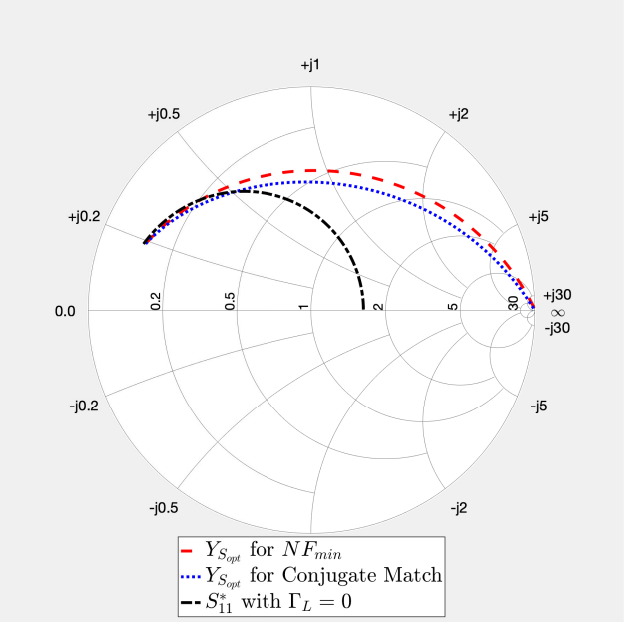


Figure 3.22: Optimum source reflection in different cases

Compared to the input impedance for a matched load, Fig. 3.23 shows that the noise of the optimal source impedance approaches the matched condition as the frequency increases. The critical observation here is that $R_S = \frac{1}{g_m}$ is not the correct choice of transistor conductance despite the low-frequency case. Simulation is required to find the optimal conductance for a given frequency.

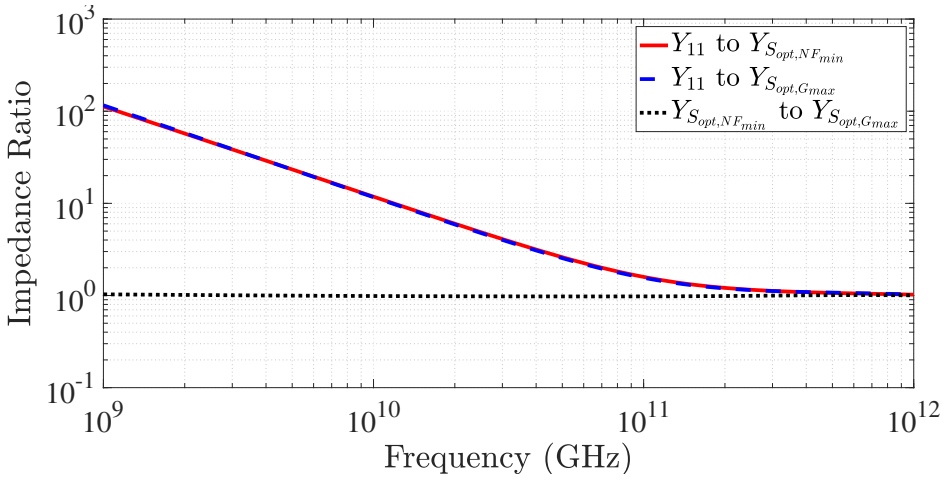


Figure 3.23: Optimum impedance ratios vs. Frequency

Once the bias circuit is included, the symmetry of the circuit is partially broken due to the body effect of the CG stage and the voltage division caused by the capacitive coupling of the CS stage. Therefore, it is important to make the layout as symmetrical as possible. Once the DC circuit is implemented, the Y parameters are extracted to find the optimal common-mode impedance. As shown in Fig. 3.25, the matching network connected to the output of the active balun should have a common mode inductance of about 400pH.

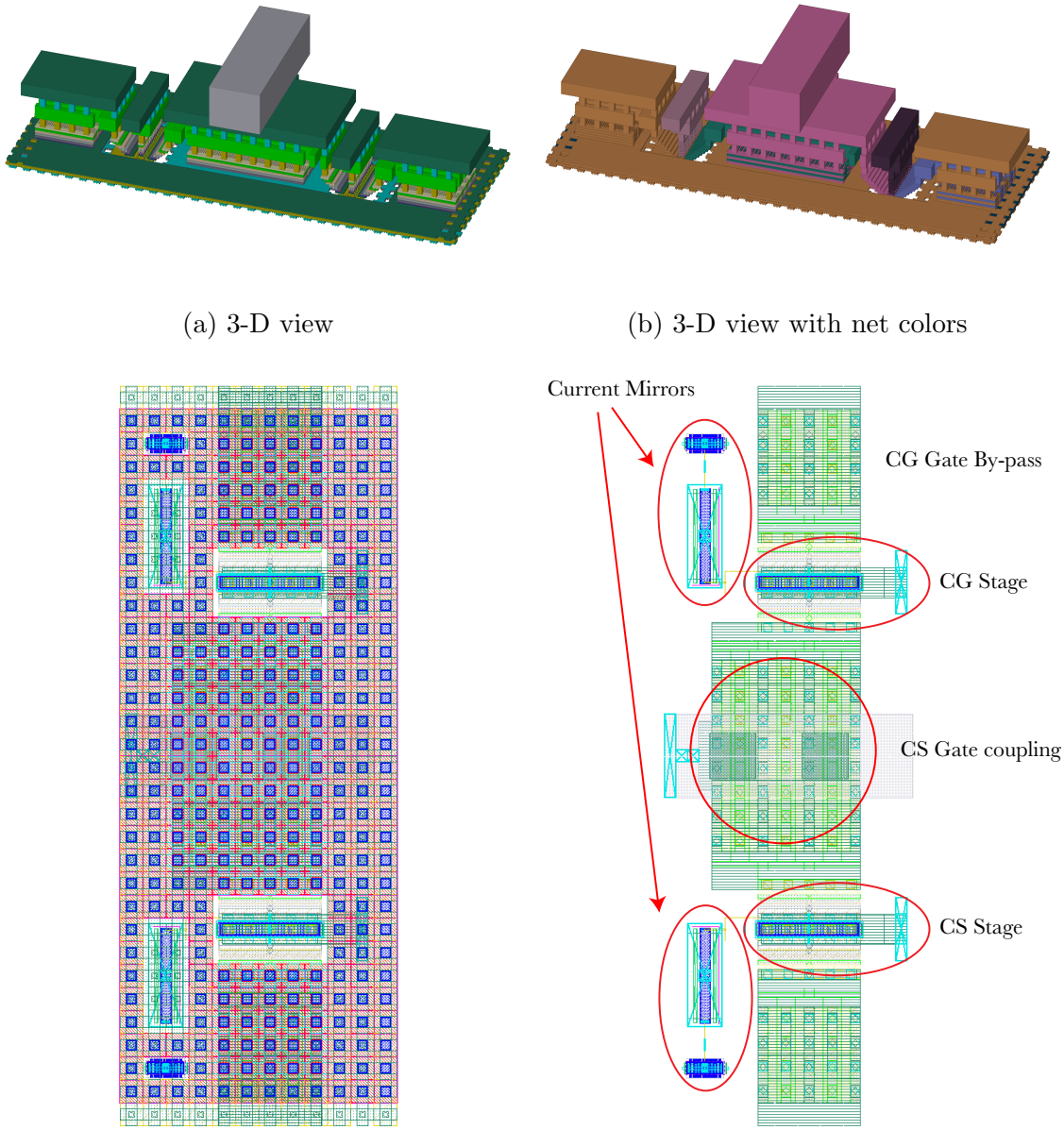
So far, the limited quality factor of the matching network has not been considered. A Python script has been written to establish a connection between the EM simulator and the circuit simulator. The performance of the LNA is calculated for several different matching networks. Fig. 3.26 shows the final implemented matching networks. Fig. 3.27 shows the input impedance of the matching network. Note that the differential inductance is relatively constant [3], while the common-mode inductance increases at about 155GHz and then goes to negative values. Note that based on Fig. 3.25, even negative values of inductance can potentially degrade the performance of the LNA.

Fig. 3.28 shows the performance of the active balun developed here. It has a power gain of 2.25dB and an insertion loss of 3.4dB due to the matching network.

3.5 Interstage Amplifiers

Once the active balun generates the differential signal, it is passed through dummy-neutralized pseudo-differential CS amplifiers shown in Fig. 3.29a. The dummy device uses high threshold voltage transistors to ensure that the channel has the highest resistance. In the capacitively

[3]Which means that the self-resonance frequency of the differential mode is high.



(a) 3-D view

(b) 3-D view with net colors

(c) Layout view

(d) Simplified layout view

Figure 3.24: Active balun transistor core.

neutralized amplifier of Fig. 3.29b, the reverse conductance can be calculated as

$$y_{12,C} = \frac{C_{gd}s}{1 + R_g(C_{gs} + C_{gd})s} - C_n s \tag{3.101}$$

$$\approx C_{gd}s\left(\left(1 - \frac{C_n}{C_{gd}}\right) - R_g(C_{gs} + C_{gd})s\right) \tag{3.102}$$

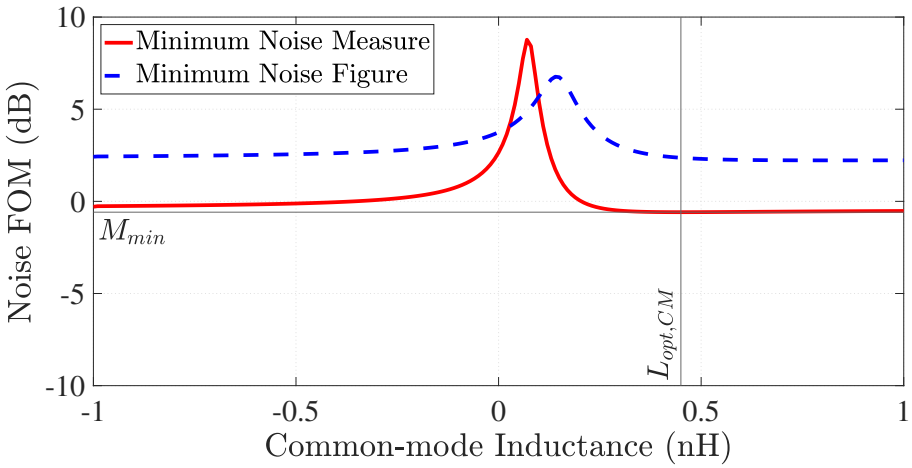
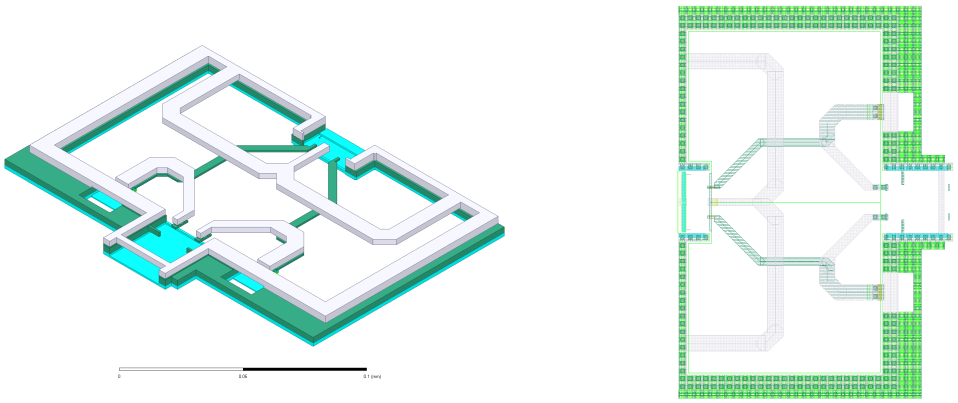


Figure 3.25: Noise performance vs. different output common-mode inductance



(a) 3-D view (b) Layout view

Figure 3.26: Matching network of the active balun.

While it is effective at low frequencies when $C_n = C_{gd}$, the reverse conductance is limited by $-C_{gd}R_g(C_{gs} + C_{gd})s^2$ as the operating frequency increases. In practice, the neutralization capacitor is made from the back-end metallization layers, while the gate-drain capacitor is from the front-end metallization layers and transistors. Since they have completely different origins, they will not follow each other in process variations. This means that while this neutralization technique is effective in simulation, it is limited in practice. Let us now

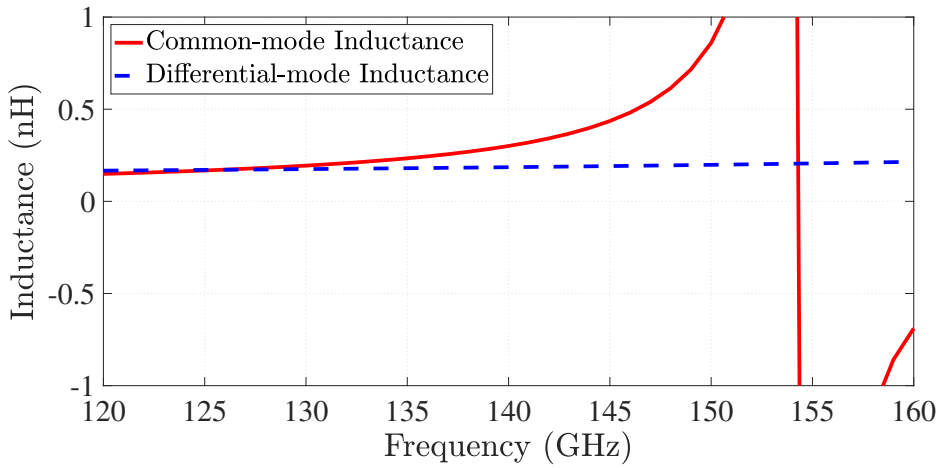


Figure 3.27: Differential and common-mode impedance of the matching network

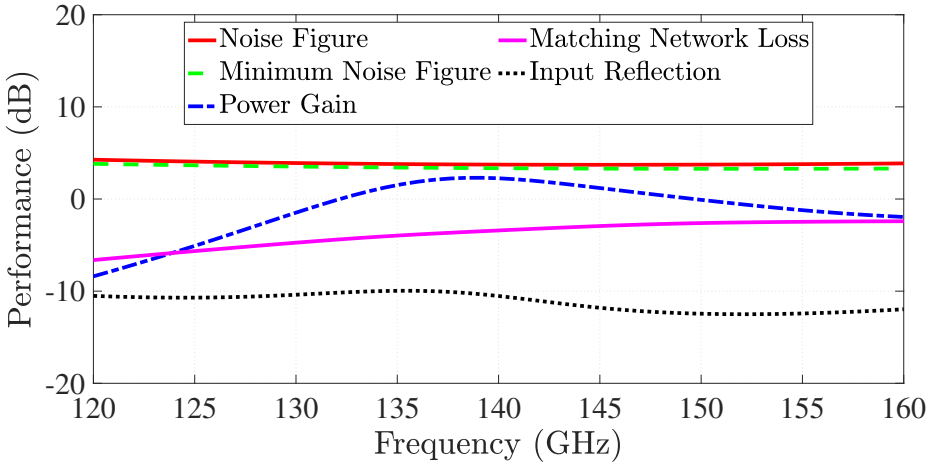


Figure 3.28: Performance of the active balun with matching network

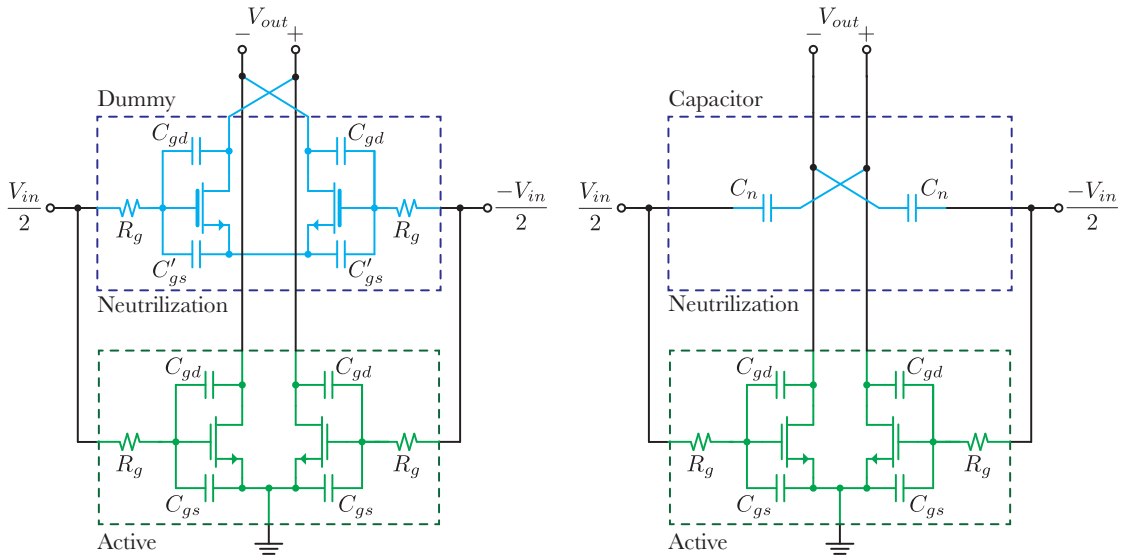
consider a dummy-neutralized topology from Fig. 3.29a with

$$y_{12,D} = \frac{C_{gd}s}{1 + R_g(C_{gs} + C_{gd})s} - \frac{C_{gd}s}{1 + R_g(C'_{gs} + C_{gd})s} \tag{3.103}$$

$$= \frac{R_g C_{gd}(C'_{gs} - C_{gs})s^2}{1 + R_g(C_{gs} + C_{gd})s} \tag{3.104}$$

$$\approx C_{gd}s\left(R_g(C'_{gs} - C_{gs})s\right)\left(1 - R_g(C_{gs} + C_{gd})s\right) \tag{3.105}$$

Note that C'_{gs} is the gate-source capacitance of the dummy device, while C_{gs} is the counterpart of the active device. Since the channel is not formed in the off device, the channel capacitance has a high series resistance that effectively blocks its action. Therefore, C'_{gs} is mainly the overlap and fringe capacitance of the front-end metallization. In the current



(a) Dummy-neutralized (b) Capacitively-neutralized

Figure 3.29: Neutralized pseudo-differential CS amplifiers

process, the simulation results show that $\frac{C_{gs}}{C'_{gs}} \approx 2$. Despite the constraint of $C'_{gs} - C_{gs}$, the dummy neutralization outperforms the capacitive neutralization topology as long as $R_g(C'_{gs} - C_{gs})s \ll 1$.

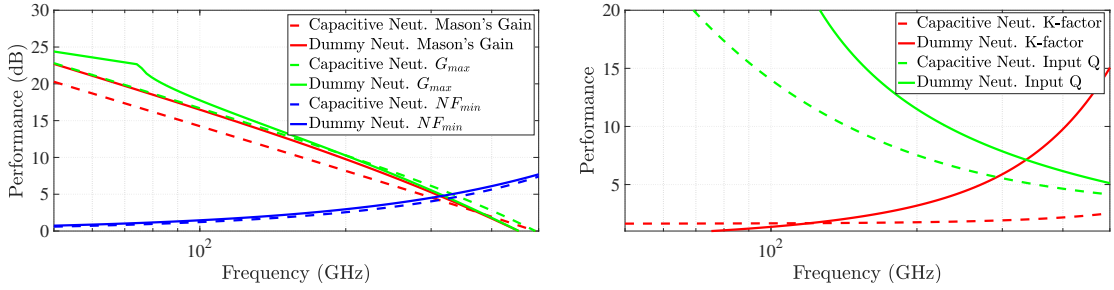
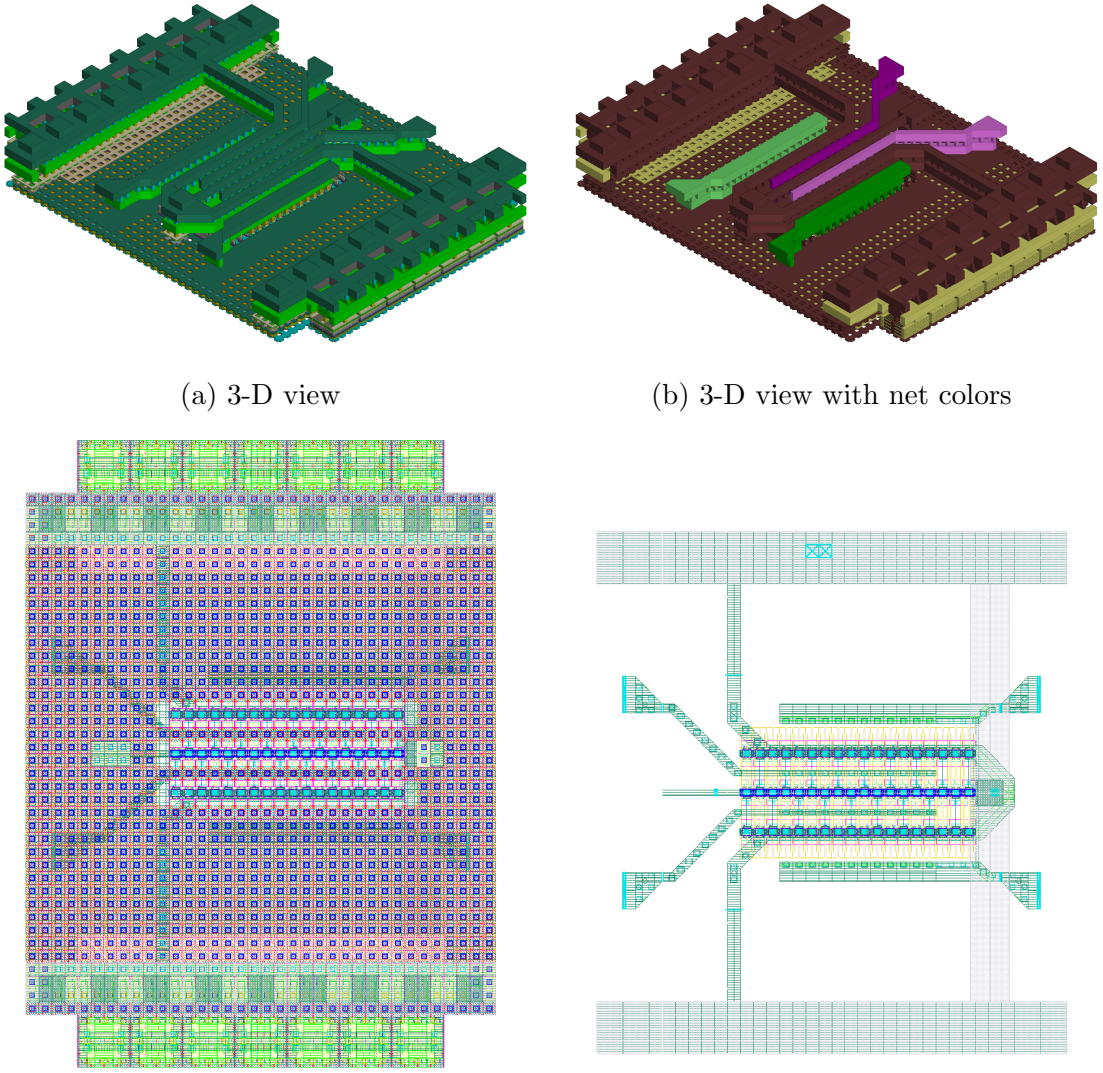


Figure 3.30: Comparison of dummy-neutralized and capacitively-neutralized amplifiers (Post layout-extraction up to M1)

Note that a lossy matching network inevitably increases the noise measure [21]. However, it can potentially increase Mason's unilateral gain [31]. As you can see in Fig. 3.32, the dummy-neutralized amplifier has a higher available gain and also a higher input quality factor [4]. The higher quality factor means that the matching network can expect a higher insertion loss.

[4]The quality factor is calculated based on the quality factor of the simultaneous input-output conjugate matching impedances.

Since dummy-neutralized amplifiers are less sensitive to process variations and accurate transistor models, it was preferred over the capacitive counterpart. Fig. 3.32 shows the



(a) 3-D view

(b) 3-D view with net colors

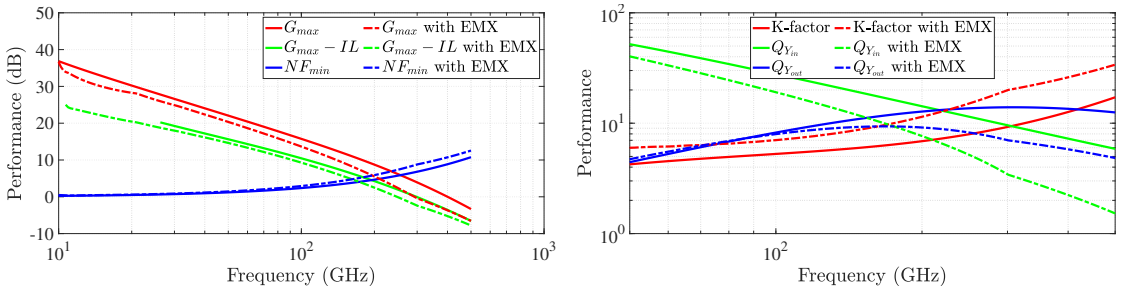
(c) Layout view

(d) Simplified layout view

Figure 3.31: Amplifier core transistors

performance of the core transistor. Note that f_{max} = 400GHz can be achieved with RC extraction. However, as soon as EMX is used, it decreases to f_{max} = 300GHz. Fig. 3.31 shows the implementation of the core transistors.

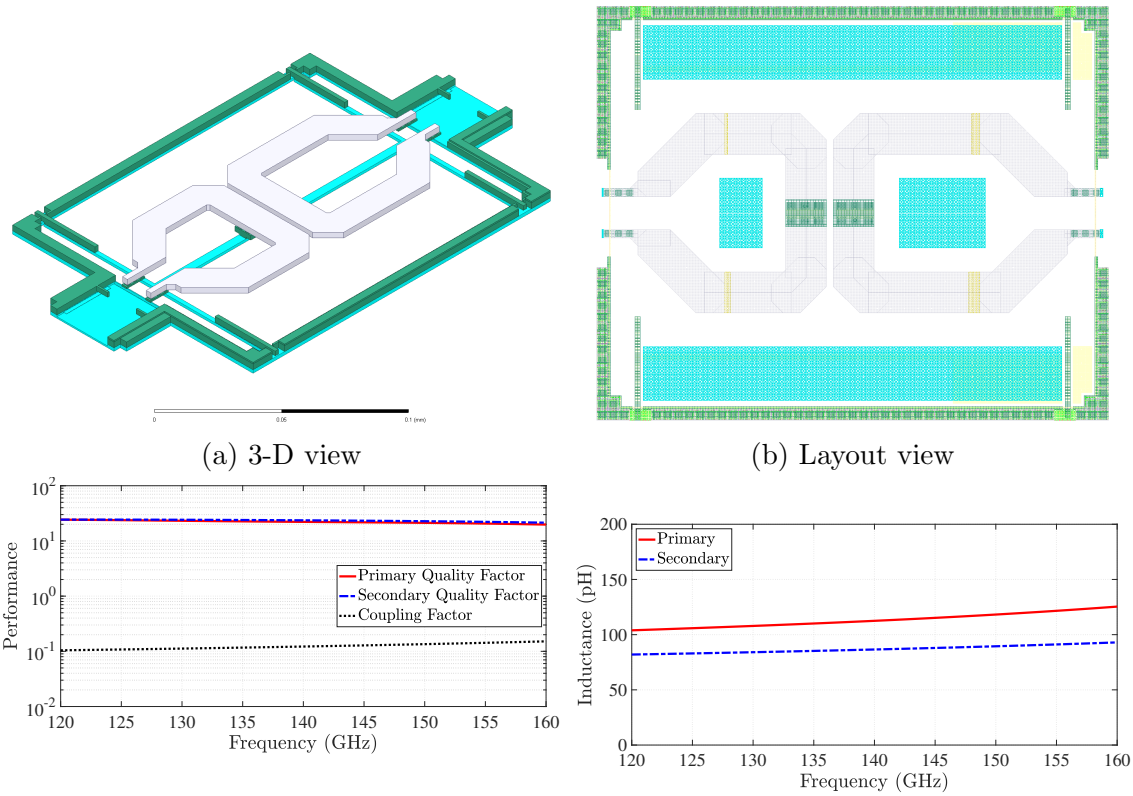
Given the high input and output quality factors, the insertion loss of the matching network must be considered. It can be estimated assuming a quality factor of 20 for the



(a) Gain and noise figure (b) Stability and termination quality factors

Figure 3.32: Performance of the amplifier core with RC-extraction and EMX

inductors of the matching network. Fig. 3.32a contains this estimate. Fig. 3.33 shows the implementation of the interstage matching network. Using the top-most thick metal, a low-k transformer with quality factors greater than 20 is implemented.



(a) 3-D view (b) Layout view

(c) Quality factor and coupling factor (d) Primary and secondary inductance

Figure 3.33: Interstage transformer

The performance of the interstage amplifier is shown in Fig. 3.34. Each amplifier consumes 4mA of DC current. Note that by using low-k transformers [32], a relatively wide

bandwidth is achieved. Reducing the coupling factor could potentially result in higher bandwidth at the expense of lower gain. However, more complicated matching networks can increase the bandwidth of the amplifier without sacrificing gain.

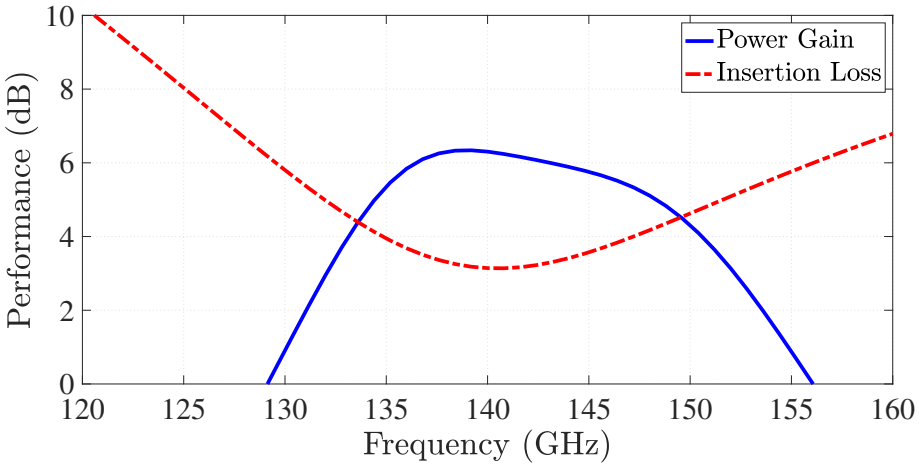
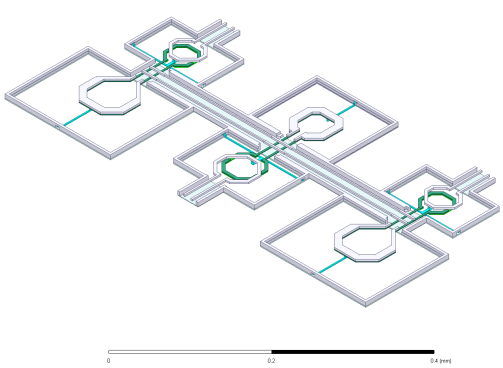


Figure 3.34: Performance of interstage amplifiers

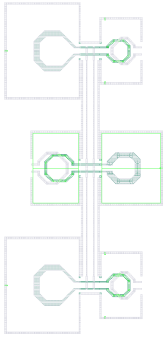
3.6 I/Q Splitter

Once the input signal is amplified, it should be split into I and Q paths before corresponding mixers. The splitter is laid out as shown in Fig. 3.35b. Note that the effective coupling factor is the same as the coupling factor of the interstage amplifier. However, an impedance transformation from 1 to 2 is realized by using transformer equivalents.

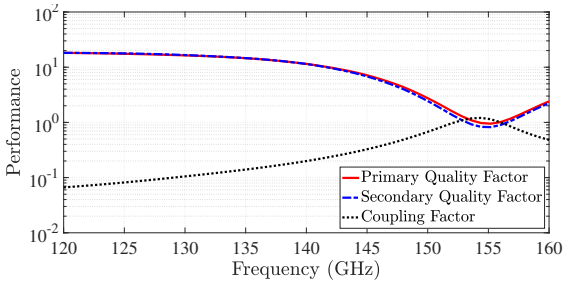
An important technique used here is the inductive termination of the transmission line. As you can see in Fig. 3.36, the optimal termination impedance for long transmission lines is Z_0 for minimum insertion loss, while small transmission lines (smaller than $\frac{\lambda}{2}$) have optimal source and load impedances, similar to the optimal source and load impedances for an amplifier. Note that the layout of the chip here determines the length of the transmission line. The exact terminations on each side of the splitter are optimized in ADS to show a wideband Chebyshev response with the effective shunt capacitance of the transmission line. Fig. 3.37 delivers the performance of the I/Q splitter. Note that as shown in Fig. 3.1, another stage of the dummy-neutralized amplifier is used after the splitter to provide isolation between the I/Q mixers.



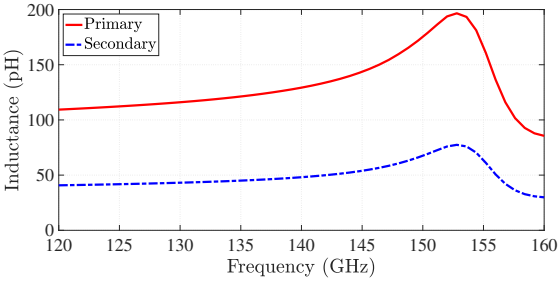
(a) 3-D view



(b) Layout view



(c) Quality factor and coupling factor



(d) Primary and secondary inductance

Figure 3.35: Splitter and its performance

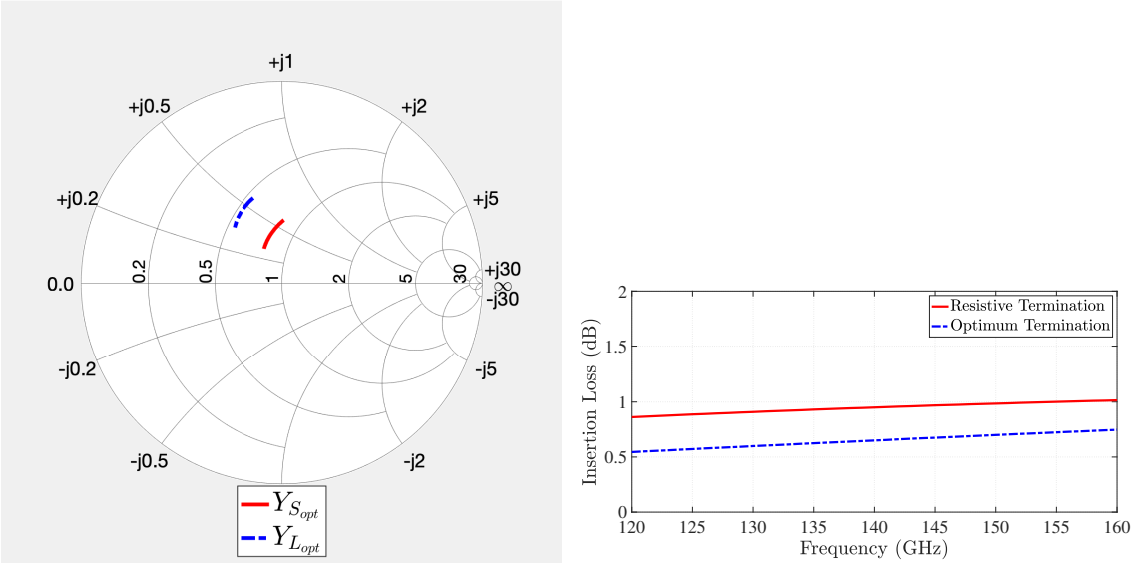
3.7 Mixer Design

In the earlier work [20], active mixers were used. The reason for this was the hypothesis that using an active mixer instead of a passive mixer may be beneficial when the operating frequency is less than half of f_{max}. It is also assumed that the conversion gain of a passive mixer is more sensitive to the LO swing [33]. However, these assumptions can be refuted as follows.

First, as the transistors in Fig. 3.39 switch between on and off states, their effective g_m falls below the peak value at which f_{max} can be reached. In other words, assuming a sharp switching behavior between on and off states in Fig. 3.39, at any time, two transistors are connected to the input RF while only one of them is active.

Second, the odd harmonic currents of the transistors do not produce a noticeable voltage swing when the impedance of the even harmonics at the common node of the mixer is low. Therefore, the switching behavior of active and passive mixers is similar. Quantitatively, in each transistor

$$I_{ds} \approx g_{m1}V_{gs} + g_{m2}{V_{gs}}^2 + g_{m3}{V_{gs}}^3 + ... \tag{3.106}$$



(a) Optimum source and load reflections (b) Insertion loss

Figure 3.36: Performance of a transmission line with different terminations

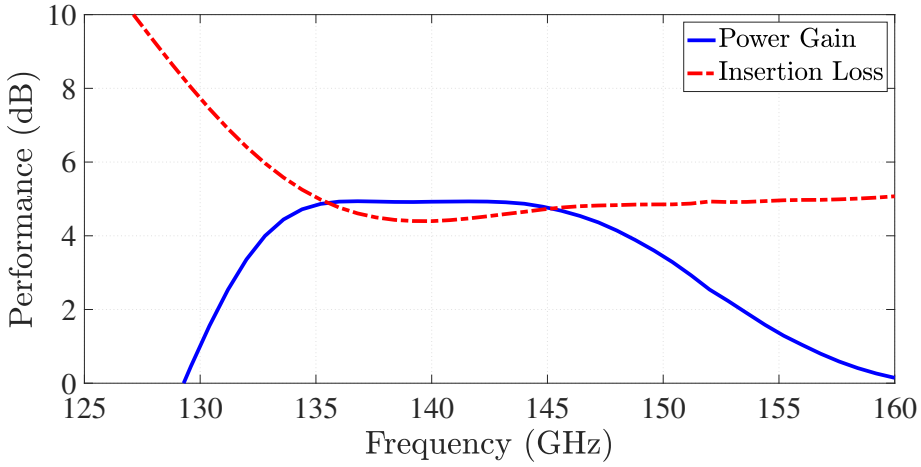


Figure 3.37: Performance of the amplifier driving the splitter with the insertion loss of the splitter

Assuming that $V_{gs} = V_{LO}\cos(\omega_{LO}t) + V_{in}\cos(\omega_{in}t + \phi)$, the current harmonics are generated at $m\omega_{LO} + n\omega_{in}$. These harmonics are passed through Z_s and change the source voltage. When $Z_s \approx 0\Omega$ for all these harmonics, the source voltage remains constant. This is the case for most millimeter-wave mixers beyond $\frac{f_t}{2}$. Assume that the source is tuned to the

fundamental frequency of ω_{LO},

$$\frac{V_s(\omega_{LO})}{V_s(2\omega_{LO})} = \left|\frac{I_{ds}(\omega_{LO})}{I_{ds}(2\omega_{LO})}\right| \left|\frac{Z_s(\omega_{LO})}{Z_s(2\omega_{LO})}\right| \tag{3.107}$$

$$\approx \left|\frac{I_{ds}(\omega_{LO})}{I_{ds}(2\omega_{LO})}\right| \frac{Q_M \frac{1}{\omega_{LO}C_s}}{\frac{1}{2\omega_{LO}C_s}} \tag{3.108}$$

$$\approx \left|\frac{I_{ds}(\omega_{LO})}{I_{ds}(2\omega_{LO})}\right| 2Q_M \tag{3.109}$$

where Q_M is the quality factor of the matching network and transistors at the fundamental frequency. Note that for devices with weak nonlinearity, the first term in the above equation is larger than 1. Therefore, despite the existence of harmonic currents, the harmonic voltages on the source side are negligible. This suggests that mixers should be operated with voltage sources rather than current sources when performing simulations to gain insight into the design space.

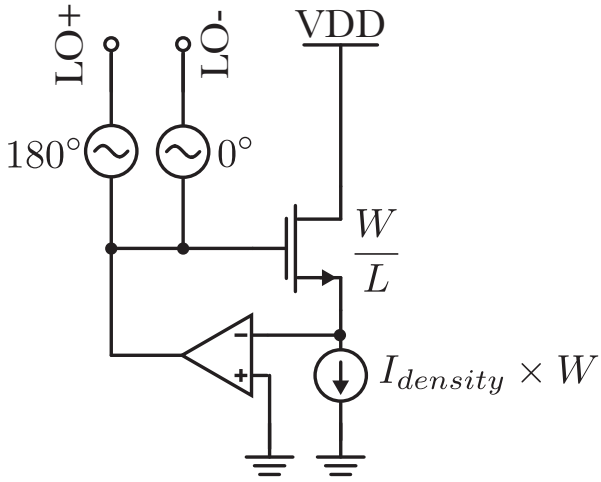


Figure 3.38: Bias generation circuit for mixers

In the rest of this section, the mixer bias is illustrated by Fig. 3.38. In the active mixer, the drain nodes are connected to the supply through ideal current sources, while in the passive mixer, the drain nodes are disconnected from the supply to be biased in the triode region. This ensures that the only difference between the two mixers in the simulation environment is the DC voltage of the drain nodes.

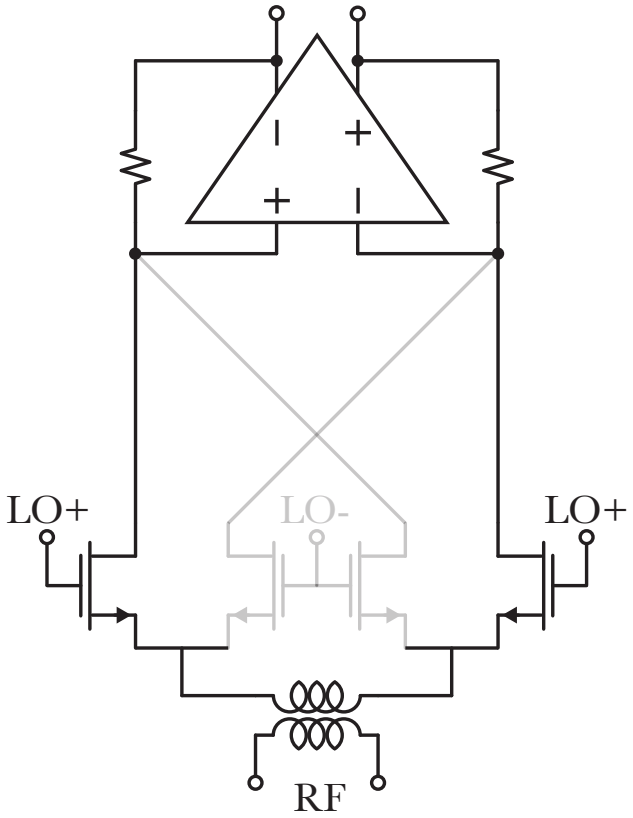


Figure 3.39: Current mode mixer

Current Mode Mixer

The current conversion efficiency can be defined as the current delivered to the load normalized by the real part of the current generated by the LNA

$$\eta = \frac{|I_{out}|}{\Re(I_{in})} \tag{3.110}$$

Note that while the passive and active mixers have relatively similar performance, as shown in Fig. 3.41a and Fig. 3.41b, high transconductance is required for the TIA to avoid voltage division in the passive mixer. On the other hand, if active mixers are used, the flicker noise of the mixer adds directly to the output, which can degrade the noise figure.

Voltage Mode Mixer

Assume a square-law device in the triode region,

$$I_{ds} = k' \left((V_{gs} - V_{th}) - \frac{V_{ds}}{2} \right) V_{ds} \tag{3.111}$$

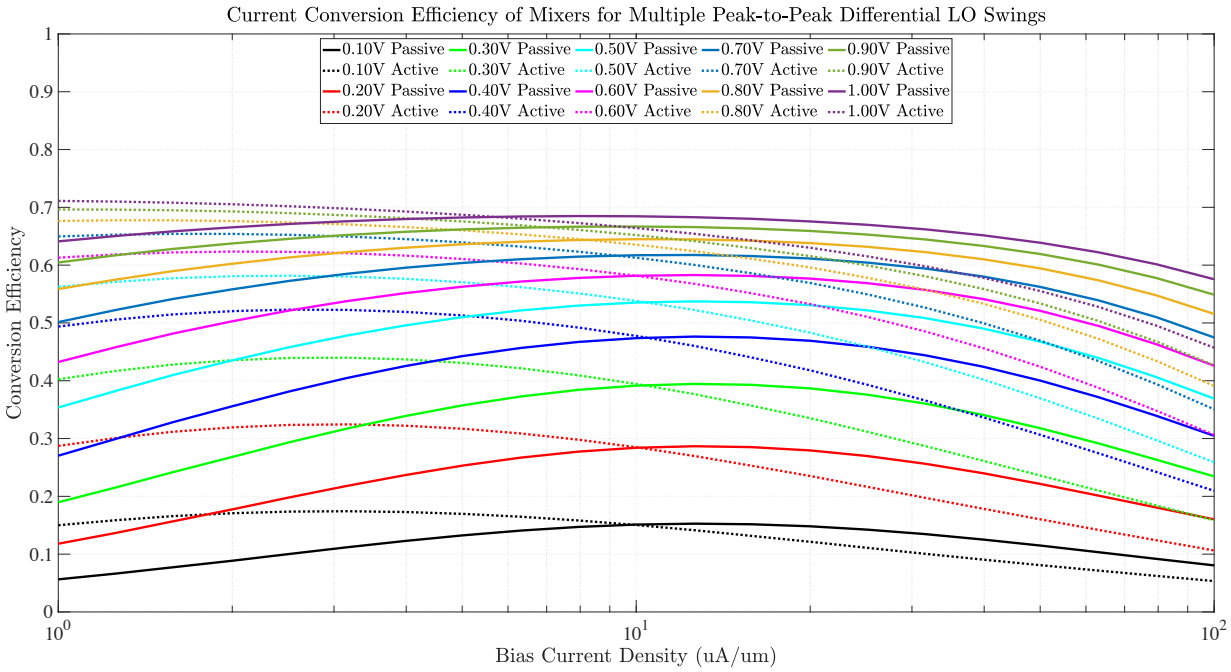
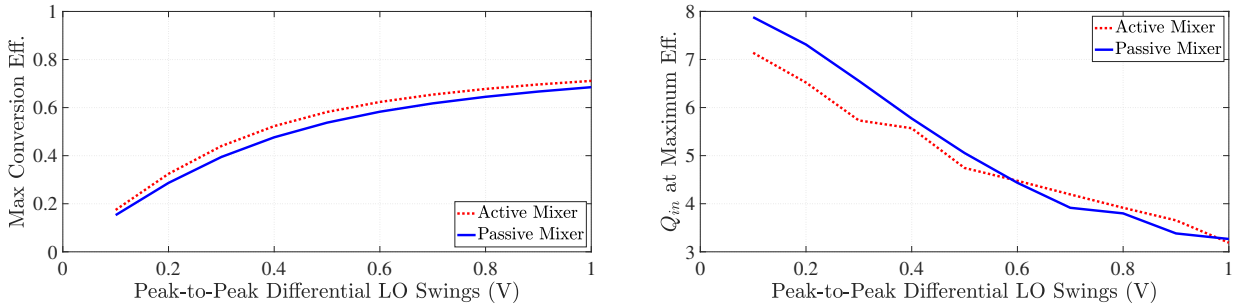


Figure 3.40: Current efficiency

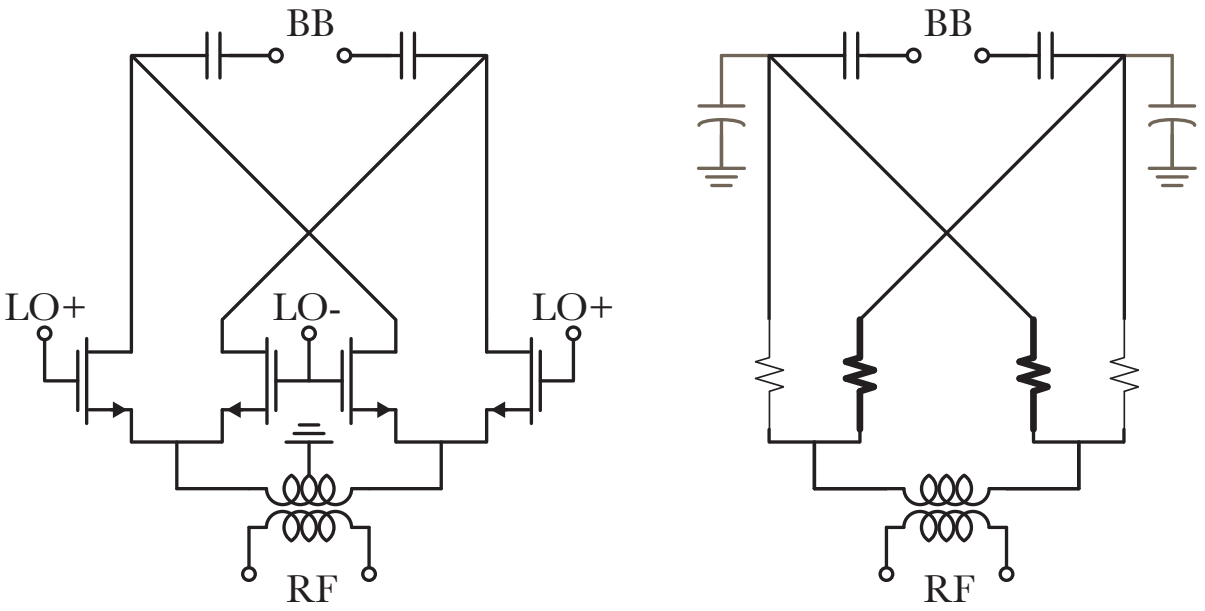


(a) Conversion efficiency (b) Input quality factor

Figure 3.41: Comparison of peak current conversion efficiency and corresponding input quality factor

and therefore, assuming a small signal variation in the drain-source voltage.

$$R_{ds} \approx \frac{1}{k' \left(V_{gs} - V_{th} \right)} \tag{3.112}$$



(a) Schematic of the mixer (b) Gain model for passive mixers

Figure 3.42: Voltage mode mixer

As for the voltage division, assuming that the LO swing is smaller than the threshold voltage ($V_{gs} > V_{th}$)

$$G_{on} = \frac{R_{ds,off}}{R_{ds,on} + R_{ds,off}} \tag{3.113}$$

$$= \frac{V_{od} + V_{LO}}{(V_{od} + V_{LO}) + (V_{od} - V_{LO})} \tag{3.114}$$

$$= \frac{1}{2}\left(1 + \frac{V_{LO}}{V_{od}}\right) \tag{3.115}$$

Similarly, the voltage division in the off-state is

$$G_{off} = \frac{1}{2}\left(1 - \frac{V_{LO}}{V_{od}}\right) \tag{3.116}$$

Assuming a sharp LO swing, when the LO signal is high

$$V_{out} = V_{in} \times (G_{on} - G_{off}) \tag{3.117}$$

$$= V_{in} \times \frac{V_{LO}}{V_{od}} \tag{3.118}$$

and when the LO signal is low

$$V_{out} = V_{in} \times (G_{off} - G_{on}) \tag{3.119}$$

$$= -V_{in} \times \frac{V_{LO}}{V_{od}} \tag{3.120}$$

which means that the input signal $V_{in}(t)$ is multiplied by $\frac{V_{LO}}{V_{od}}\omega_{LO}\sin(\omega_{LO}t)$. Note that increasing the overdrive voltage decreases the conversion gain. The maximum gain can be reached when the LO swing magnitude reaches the overdrive voltage. At higher LO swings, the off switch has a very low conductance, and the gain remains constant.

Note that when V_{od} is decreased, the bandwidth decreases despite the improvement in conversion gain. This is because the on-resistance of the switch increases, reducing the ability of the switch to drive load capacitors. So there is a tradeoff between the gain and bandwidth of mixers.

It should be noted that the maximum conversion gain of the passive mixer can be higher than $\frac{2}{\pi}$ because as the LO swing increases beyond the overdrive voltage, the conduction angle decreases, and the mixer becomes more similar to a sample-and-hold circuit. If $V_{LO} \leq V_{od}$, the equivalent Thevenin voltage source is

$$V_{out}(t) = V_{in}(t) \times \frac{V_{LO}}{V_{od}}\sin(\omega_{LO}t) \tag{3.121}$$

If $V_{LO} \gg V_{od}$, the Thevenin equivalent voltage source can be approximated as

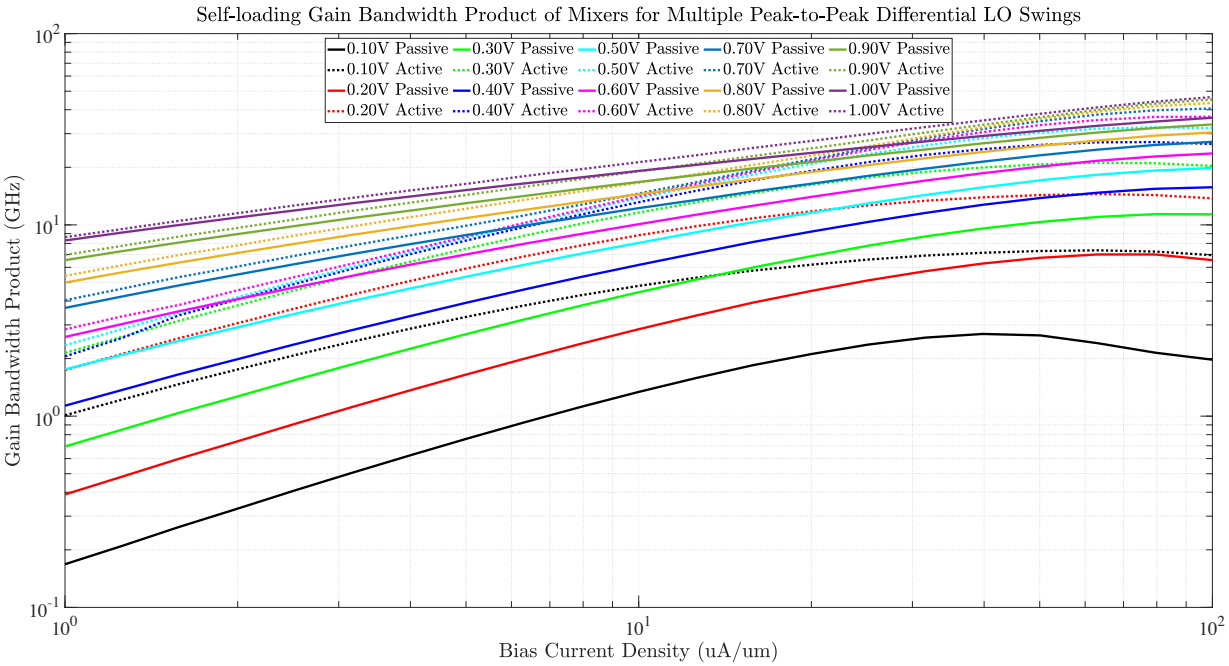
$$V_{out}(t) \approx V_{in}(t)\Pi(\omega_{LO}t) \tag{3.122}$$

However, in the presence of the sampling capacitor, the resistance of the Thevenin equivalent source charging the sampling capacitor should also be considered. As shown in Fig. 3.44, if the $V_{od} \to 0$, the conduction time of the Thevenin equivalent resistor decreases. Therefore, the modulated input signal is first sampled and then held. This additional sampling mechanism downconverts the upconverted spectral content of the signal and increases the theoretical conversion gain of the passive mixer to 0dB.

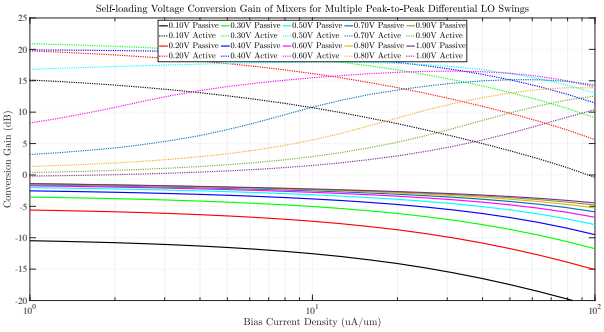
As you can see in Fig. 3.43a, the gain-bandwidth product of the active and passive mixers in the voltage mode is quite similar. Assuming that the mixer has less than 1dB attenuation at the edge of the desired bandwidth, the 3dB bandwidth should be twice the desired baseband bandwidth since

$$\frac{1}{1+\left(\frac{f}{f_{3\text{dB}}}\right)^2} = 10^{\frac{-1}{10}} \Rightarrow f \approx \frac{f_{3\text{dB}}}{2} \tag{3.123}$$

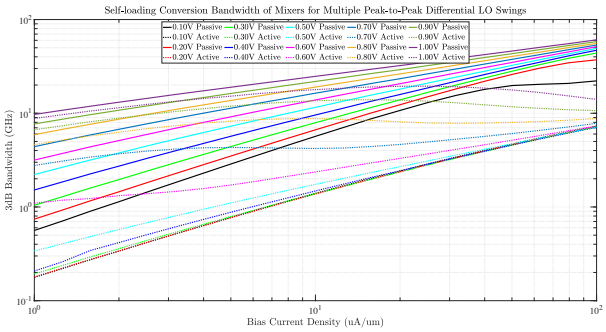
A passive mixer is used to reduce the power consumption of the array elements. It also has lower flicker noise compared to its active counterpart.



(a) Gain-Bandwidth Product



(b) Conversion Gain



(c) −3dB Bandwidth

Figure 3.43: Comparison of active and passive mixers in voltage mode with different peak-to-peak differential LO swings

Since passive mixers are reciprocal, the input impedance of the mixer should be investigated. [34] provides an excellent mathematical framework for calculating the input impedance. The current can be calculated as

$$I_{in}(t) = \Pi(\omega_{LO}t)S(2\omega_{LO}t)\left[V_{in}(t)\Pi(\omega_{LO}t) * \mathfrak{f}\left\{Y_L(t), S(2\omega_{LO}\tau)\right\}\right] \tag{3.124}$$

where $\mathfrak{f}\left\{Y_L(t), S(2\omega_{LO}\tau)\right\}$ is the current response of the system at time t to an impulse voltage at time τ.

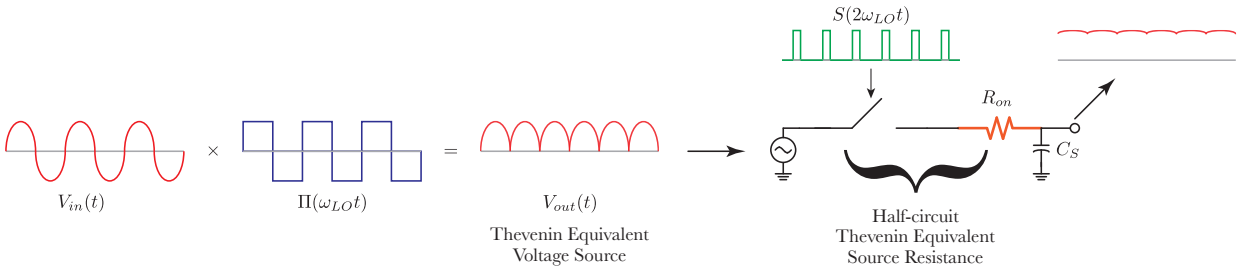


Figure 3.44: Equivalent Thevenin source used in the mixer model

$$\mathfrak{f}\{Y_L(t), S(2\omega_{LO}\tau)\} = \begin{cases} 0, & \text{if } S(2\omega_{LO}\tau) = 0 \\ \frac{\delta(t)}{R_{on}} - \frac{e^{\frac{-\int_0^t S(2\omega_{LO}x)dx}{R_{on}C_S}}}{{R_{on}}^2 C_S}, & \text{if } S(2\omega_{LO}\tau) = 1 \end{cases} \tag{3.125}$$

By algebraic manipulation, the input current can be written as

$$I_{in}(t) = \Pi(\omega_{LO}t)S(2\omega_{LO}t)\left[V_{in}(t)\Pi(\omega_{LO}t)S(2\omega_{LO}t) * Y_L'(t)\right] \tag{3.126}$$

where

$$Y_L'(t) = \frac{\delta(t)}{R_{on}} - \frac{e^{\frac{-\int_0^t S(2\omega_{LO}x)dx}{R_{on}C_S}}}{{R_{on}}^2 C_S} \tag{3.127}$$

Approximating the integral part in the exponential decay with its continuous-time equivalent simplifies the above equation into

$$Y_L'(t) \approx \frac{\delta(t)}{R_{on}} - \frac{e^{\frac{-\alpha t}{R_{on}C_S}}}{{R_{on}}^2 C_S} \tag{3.128}$$

where α is the mean term of $S(2\omega_{LO}t)$. In the Laplace domain, $Y_L'(s)$ can be written as

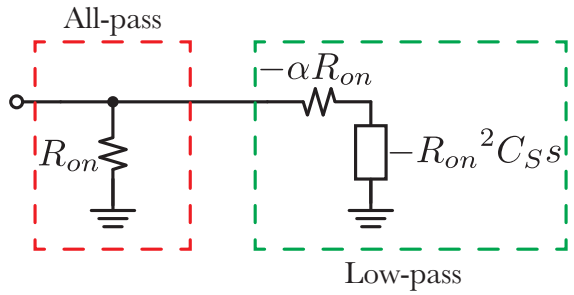


Figure 3.45: Decomposition of the impedance seen by the equivalent source into an all-pass and a low-pass section

$$Y_L'(s) = \left[R_{on} || \left(-\alpha R_{on} - {R_{on}}^2 C_S s\right)\right]^{-1} \tag{3.129}$$

which can be divided into an all-pass and a low-pass section, as in Fig. 3.45. Since Eq. 3.126 is a linear equation, $I_{in}(t)$ can be calculated as the response to each section. Note that the response to the all-pass section can be calculated simply as

$$I_{in,\text{all-pass}}(t) = \Pi(\omega_{LO}t)S(2\omega_{LO}t)\left[V_{in}(t)\Pi(\omega_{LO}t)S(2\omega_{LO}t) * \frac{\delta(t)}{R_{on}}\right] \tag{3.130}$$

$$= \frac{V_{in}(t)}{R_{on}}S(2\omega_{LO}t)^2 \tag{3.131}$$

$$\approx \frac{\alpha^2 V_{in}(t)}{R_{on}} \tag{3.132}$$

where the last approximation ignores the harmonics of $S(2\omega_{LO}t)$. Computing the response to the low-pass section is more complicated. In the Laplace domain

$$I_{in,\text{low-pass}}(s) = \Pi(s) * S(s) * \left[(V_{in}(s) * \Pi(s) * S(s)) \times Y'_{L,\text{low-pass}}(s)\right] \tag{3.133}$$

Assuming that the high-frequency current of the low-pass is negligible [5],

$$I_{in,\text{low-pass}}(s) \approx \left(\frac{2}{\pi}\right)^2 \alpha^2 V_{in}(s) Y'_{L,\text{low-pass}}(s') \tag{3.134}$$

where $s' = j|(\omega - \omega_{LO})|$. Note that for the frequency range outside the baseband bandwidth,

$$R_{in,\text{out-of-band}} \approx \alpha^2 R_{on} \tag{3.135}$$

and for the frequency range within the baseband bandwidth,

$$R_{in,\text{in-band}} \approx \alpha^2 \frac{R_{on}}{1 - \left(\frac{2}{\pi}\right)^2 \frac{1}{\alpha}} \tag{3.136}$$

Fig. 3.46 shows the ratio between the in-band input resistance and the out-of-band input resistance. Note that at low LO swings, the assumption of hard switching of the mixer does not hold, and the above model breaks. Both the on and off switches are somewhat conductive in this case, resulting in dissipative behavior with no conversion gain. As the LO swing continues to increase, the hard switching becomes more realistic, and with $\alpha \approx 1$ exactly at the conversion gain of $\frac{2}{\pi}$,

$$\frac{R_{in,\text{in-band}}}{R_{in,\text{out-of-band}}} \approx \frac{1}{1 - \left(\frac{2}{\pi}\right)^2} \tag{3.137}$$

As the LO swing continues to increase, R_{on} decreases. However, this lower value of R_{on} is reached for a shorter time, i.e., $\alpha < 1$. Thus, the ratio of the two resistors increases. Therefore, despite its existence, the baseband capacitor is not visible in the RF domain, and

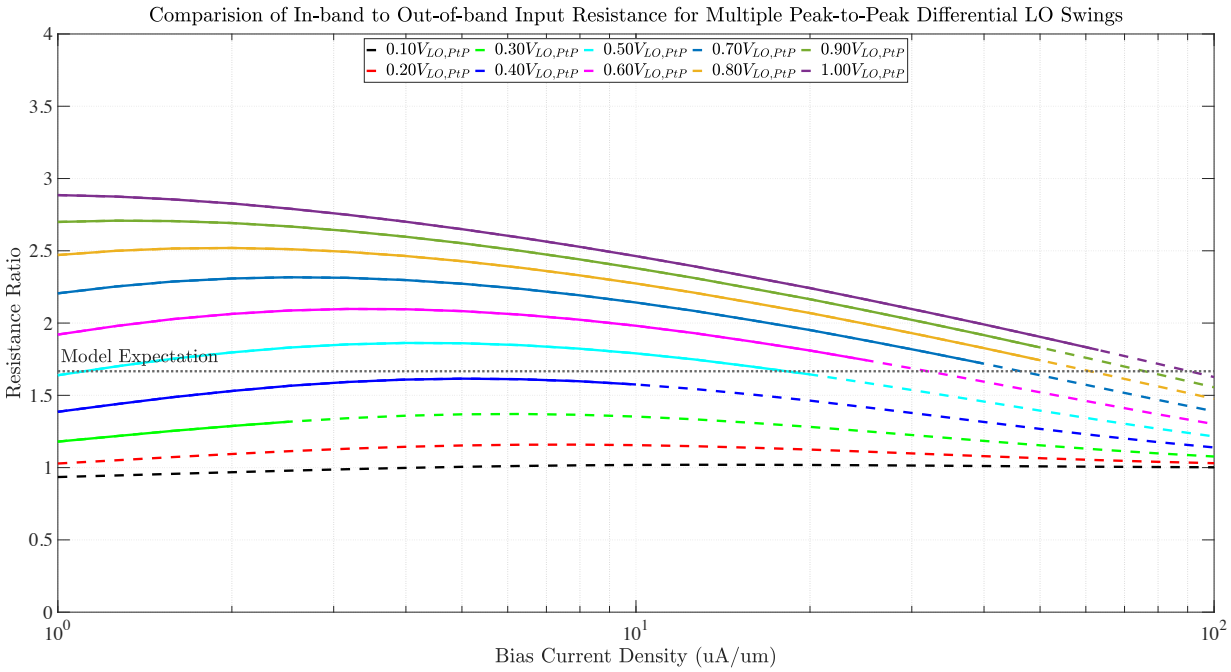
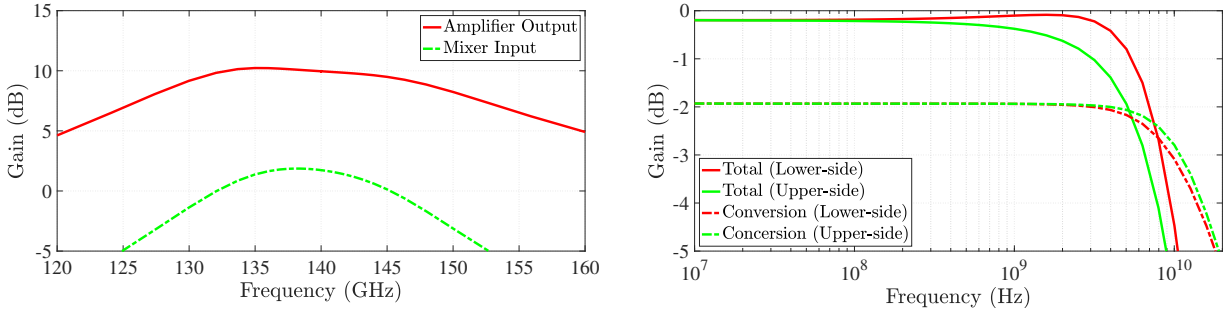


Figure 3.46: Comparison of the input resistance of the passive mixers for in-band and out-of-band tones. The dashed portion of each line shows the region where the gain falls below $\frac{2}{\pi}$.



(a) Millimeter- wave gain (b) Conversion Gain

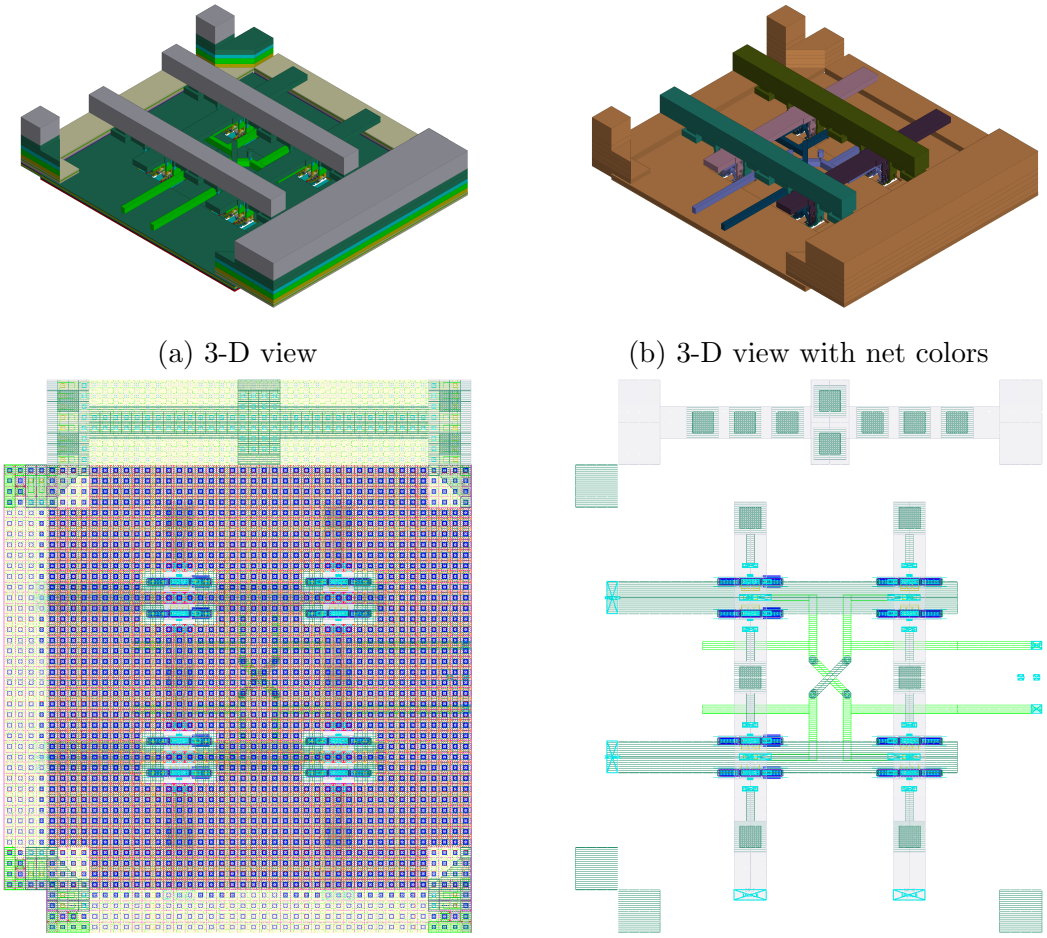
Figure 3.47: Performance of the mixer and its preceding gain stage

the passive mixer exhibits an input impedance with a relatively low quality factor. Note that the gate capacitance and the parasitic elements of the layout increase the quality factor.

Fig. 3.47b shows the performance of the mixer when driven with 660mV LO swing. The DC bias of switches is generated by a current mirror biased at a current density of $100\mu A/\mu m$. While the conversion gain itself has a high bandwidth, the overall conversion

[5]This requires that the baseband bandwidth to be much smaller than the carrier frequency, a condition that does not hold for wideband communication links.

gain has a limited bandwidth of 2×7GHz due to an error in the matching network of the previous buffer stage. The problem with the matching network is illustrated in Fig. 3.47a. While the AC voltage gain at the output of the amplifier is broadband, the low coupling factor of the transformer reduces the overall bandwidth at the input of the mixer. Note that a 3dB attenuation budget cannot be used exclusively in the mixer since it is cascaded with the rest of the chain. Fortunately, my colleague Ethan Chou caught this error on the second tapeout and corrected it.



(a) 3-D view

(b) 3-D view with net colors

(c) Layout view

(d) Simplified layout view

Figure 3.48: Mixer implementation

3.8 Baseband Amplifier

The previous millimeter-wave transceiver used a Cherry-Hooper amplifier [35] (Fig. 3.49). Despite its broadband performance and high gain, it had some problems:

- The amplifiers were designed as pseudo-differential stages. This topology is prone to common-mode noise since any unwanted coupling (from surrounding circuits or the supply network) goes through a high-gain amplification chain. Although the output is differential, the common-mode noise may saturate the intermediate blocks, resulting in a low differential gain.
- Since the amplifiers themselves are self-biassed, the current consumption of each stage is highly process-dependent.

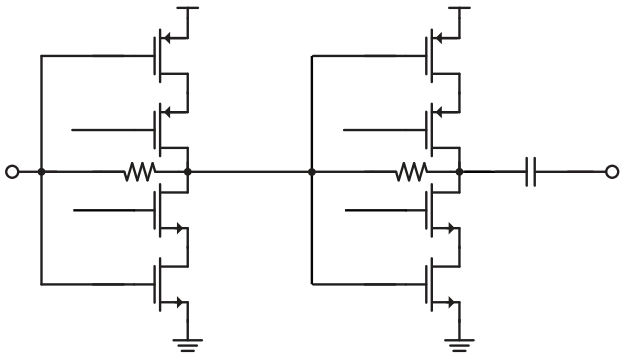


Figure 3.49: Wideband Cherry-Hooper amplifier [35]

It should be noted that the implemented Cherry-Hooper amplifier still consists of multiple cascaded amplifiers. Therefore, it is instructive to investigate why and how a Cherry-Hooper outperforms cascaded amplifiers. Note that if each stage has a simple first-order frequency response ([36])

$$A(s) = \frac{A_0}{1 + \frac{s}{\omega_0}} \tag{3.138}$$

the bandwidth for α attenuation is

$$1 + \left(\frac{\omega}{\omega_0}\right)^2 = \alpha^{-1} \Rightarrow BW = \omega_0\sqrt{\alpha^{-1} - 1} \tag{3.139}$$

and therefore the gain-bandwidth product $GBW_0 = A_0\omega_0\sqrt{\alpha^{-1} - 1}$. By cascading N of these stages one obtains

$$A_N(s) = \frac{{A_0}^N}{\left(1 + \frac{s}{\omega_0}\right)^N} \tag{3.140}$$

which corresponds to a bandwidth of

$$\left(1 + \left(\frac{\omega}{\omega_0}\right)^2\right)^N = \alpha^{-1} \Rightarrow BW_N = \omega_0\sqrt{\alpha^{\frac{-1}{N}} - 1} \tag{3.141}$$

and thus $GBW_N = {A_0}^N \omega_0 \sqrt{\alpha^{\frac{-1}{N}} - 1}$. The gain expansion can be calculated as

$$\frac{GBW_N}{GBW_0} = \frac{{A_0}^N \omega_0 \sqrt{\alpha^{\frac{-1}{N}} - 1}}{A_0 \omega_0 \sqrt{\alpha^{-1} - 1}} \tag{3.142}$$

$$= \left({A_0}^N\right)^{1-\frac{1}{N}} \frac{\sqrt{\alpha^{\frac{-1}{N}} - 1}}{\sqrt{\alpha^{-1} - 1}} \tag{3.143}$$

$$= {A_{tot}}^{1-\frac{1}{N}} \frac{\sqrt{\alpha^{\frac{-1}{N}} - 1}}{\sqrt{\alpha^{-1} - 1}} \tag{3.144}$$

Note that the maximum bandwidth expansion occurs at

$$\frac{\partial \frac{GBW_N}{GBW_0}}{\partial N} = 0 \Rightarrow N_{opt} = \frac{-\ln(\alpha)}{\ln(\frac{2\ln(A_{tot})}{2\ln(A_{tot})+\ln(\alpha)})} \approx 2\ln(A_{tot}) \tag{3.145}$$

where the last approximation works at a high total gain and less than 3dB attenuation. Given the optimal number of stages, the optimal gain per stage can be easily calculated as $A_{opt} = \sqrt{e}$.

Now assume that each stage has a maximally flat M-th order Butterworth frequency response,

$$|A(\omega)| = \frac{A_0}{\sqrt{1 + \left(\frac{\omega}{\omega_0}\right)^{2M}}} \tag{3.146}$$

and its bandwidth for α attenuation is defined as

$$1 + \left(\frac{\omega}{\omega_0}\right)^{2M} = \alpha^{-1} \Rightarrow BW = \omega_0 \sqrt[2M]{\alpha^{-1} - 1} \tag{3.147}$$

and thus $GBW_0 = A_0 \omega_0 \sqrt[2M]{\alpha^{-1} - 1}$. Cascading the same N of such amplifiers results in a new amplifier with a frequency response of

$$|A_N(\omega)| = \frac{{A_0}^N}{\left(\sqrt{1 + \left(\frac{\omega}{\omega_0}\right)^{2M}}\right)^N} \tag{3.148}$$

The bandwidth of the new amplifier can be defined as

$$\left(1 + \left(\frac{\omega}{\omega_0}\right)^{2M}\right)^N = \alpha^{-1} \Rightarrow BW_N = \omega_0 \sqrt[2M]{\alpha^{\frac{-1}{N}} - 1} \tag{3.149}$$

As before, the gain-bandwidth expansion can be defined as

$$\frac{GBW_N}{GBW_0} = {A_{tot}}^{1-\frac{1}{N}} \frac{\sqrt[2M]{\alpha^{\frac{-1}{N}} - 1}}{\sqrt[2M]{\alpha^{-1} - 1}} \tag{3.150}$$

which peaks at

$$\frac{\partial \frac{GBW_N}{GBW_0}}{\partial N} = 0 \Rightarrow N_{opt} = \frac{-\ln(\alpha)}{\ln(\frac{2M\ln(A_{tot})}{2M\ln(A_{tot})+\ln(\alpha)})} \approx 2M\ln(A_{tot}) \tag{3.151}$$

and the optimal gain per stage is $A_{M,opt} = \sqrt[2M]{e}$.

When a total fan out of F_{tot} from input to output is required, the gain bandwidth expansion can be calculated as

$$\frac{GBW_{N,F_{tot}}}{GBW_0} = {A_{tot}}^{1-\frac{1}{N}} \frac{\sqrt[2M]{\alpha^{\frac{-1}{N}}-1}}{\sqrt[2M]{\alpha^{-1}-1}} \frac{1}{{F_{tot}}^{\frac{1}{N}}} \tag{3.152}$$

where it is assumed that the natural frequency of each stage scales with $\frac{1}{{F_{tot}}^{\frac{1}{N}}}$. To find the optimal number of stages

$$\frac{\partial \frac{GBW_{N,F_{tot}}}{GBW_0}}{\partial N} = 0 \Rightarrow N_{opt} = \frac{-\ln(\alpha)}{\ln(\frac{2M\ln(A_{tot})+2M\ln(F_{tot})}{2M\ln(A_{tot})+2M\ln(F_{tot})+\ln(\alpha)})} \approx 2M\ln(A_{tot})+2M\ln(F_{tot}) \tag{3.153}$$

The optimal gain per stage and fan-out per stage can be calculated as

$$A_{opt,F_{tot}} = \left(\sqrt{e}\right)^{\frac{1}{M\left(1+\frac{\ln(F_{tot})}{\ln(A_{tot})}\right)}} \tag{3.154}$$

$$F_{opt} = \left(\sqrt{e}\right)^{\frac{1}{M\left(1+\frac{\ln(A_{tot})}{\ln(F_{tot})}\right)}} \tag{3.155}$$

Note that the low gain of each stage requires multiple stages in the optimal case. Consider the power of N stages for the gain-bandwidth expansion,

$$Power_{DC,N} \propto 1 + \left({F_{tot}}^{\frac{1}{N}}\right) + \left({F_{tot}}^{\frac{1}{N}}\right)^2 + ... + \left({F_{tot}}^{\frac{1}{N}}\right)^{N-1} \tag{3.156}$$

$$\propto \frac{F_{tot}-1}{{F_{tot}}^{\frac{1}{N}}-1} \tag{3.157}$$

Therefore, we can define the efficiency of the expansion as

$$\eta = \frac{GBW_{N,F_{tot}}}{GBW_0}\frac{Power_{DC,1}}{Power_{DC,N}} = {A_{tot}}^{1-\frac{1}{N}} \frac{\sqrt[2M]{\alpha^{\frac{-1}{N}}-1}}{\sqrt[2M]{\alpha^{-1}-1}} \frac{1}{{F_{tot}}^{\frac{1}{N}}} \frac{{F_{tot}}^{\frac{1}{N}}-1}{F_{tot}-1} \tag{3.158}$$

Assuming a low fan-out per stage and a high number of stages

$$Power_{DC} \propto \frac{F_{tot}-1}{\ln(F_{tot})}N \tag{3.159}$$

which means that

$$\eta \approx {A_{tot}}^{1-\frac{1}{N}} \frac{\sqrt[2M]{\alpha^{\frac{-1}{N}}-1}}{\sqrt[2M]{\alpha^{-1}-1}} \frac{1}{{F_{tot}}^{\frac{1}{N}}} \frac{\ln(F_{tot})}{F_{tot}-1} \frac{1}{N} \tag{3.160}$$

To find the optimal efficiency,

$$\frac{\partial \eta}{\partial N} = 0 \Rightarrow 2\ln(A_{tot})M + 2M\ln(F_{tot}) = 2MN_{opt} + \frac{\ln(\alpha)}{\alpha^{\frac{1}{N_{opt}}}-1} \tag{3.161}$$

To simplify the answer, note that

$$\lim_{\alpha \to 1} \frac{\ln(\alpha)}{\alpha^{\frac{1}{N_{opt}}}-1} = N_{opt} \tag{3.162}$$

and therefore,

$$N_{opt,Power} \approx \frac{2\ln(A_{tot})M + 2M\ln(F_{tot})}{2M+1} \tag{3.163}$$

$$A_{opt,Power,F_{tot}} = e^{\frac{M+\frac{1}{2}}{M\left(1+\frac{\ln(F_{tot})}{\ln(A_{tot})}\right)}} \tag{3.164}$$

$$F_{opt,Power} = e^{\frac{M+\frac{1}{2}}{M\left(1+\frac{\ln(A_{tot})}{\ln(F_{tot})}\right)}} \tag{3.165}$$

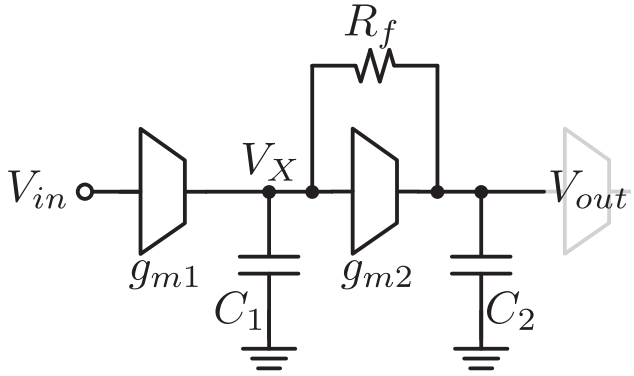


Figure 3.50: Simplified model of the Cherry-Hooper topology

Let us now analyze a simple Cherry-Hooper design from Fig. 3.50

$$V_{out}C_2 s + g_{m2}V_{in} = \frac{V_X - V_{out}}{R_f} \tag{3.166}$$

$$-g_{m1}V_{in} = V_X C_1 s + \frac{V_X - V_{out}}{R_f} \tag{3.167}$$

and the gain can be calculated as

$$\frac{V_{out}}{V_{in}}(s) = \frac{g_{m1}}{g_{m2}} \frac{g_{m2}R_f - 1}{1 + \frac{C_1+C_2}{g_{m2}}s + \frac{C_1C_2R_f}{g_{m2}}s^2} \tag{3.168}$$

For a maximally flat response

$$g_m R_f = \frac{1}{2}\frac{(C_1 + C_2)^2}{C_1C_2} \tag{3.169}$$

should be satisfied. Note that the natural frequency of this Cherry-Hooper chain is

$$\omega_{C-H} = \sqrt{2}\frac{g_{m2}}{C_1 + C_2} \tag{3.170}$$

The DC gain can be calculated as

$$A_{C-H} = \frac{1}{2}\frac{{C_1}^2 + {C_2}^2}{C_1C_2}\frac{g_{m1}}{g_{m2}} \tag{3.171}$$

Assume that $\frac{g_{m1}}{g_{m2}} = \beta$ satisfies the optimal condition for the maximum gain bandwidth. In this case

$$C_1 = C_g + \beta C_d \tag{3.172}$$

$$C_2 = \beta C_g + C_d \tag{3.173}$$

where C_g and C_d are the gate capacitance and drain capacitance of a transconductance stage, respectively. Therefore,

$$A_{C-H} = \frac{1}{2}\frac{(C_g + \beta C_d)^2 + (\beta C_g + C_d)^2}{(C_g + \beta C_d)(\beta C_g + C_d)}\beta \tag{3.174}$$

Since the optimal gain of $A_{C-H} = \sqrt[4]{e}$ is very close to 1, we will first solve this equation for $A_{C-H} = 1$ and then adjust β to reach the optimal value.

$$A_{C-H}(\beta) = 1 \Rightarrow \beta = 1 \tag{3.175}$$

$$A_{C-H}(1 + \Delta\beta) \approx \left.\frac{\partial A_{C-H}(\beta)}{\partial \beta}\right|_{\beta=1} \Delta\beta \tag{3.176}$$

To avoid tedious derivations, the gain equation can be reformulated as follows.

$$A_{C-H} = \frac{1}{2}\frac{(C_g + \beta C_d)^2 + (\beta C_g + C_d)^2}{(C_g + \beta C_d)(\beta C_g + C_d)}\beta \tag{3.177}$$

$$= \beta\left(\frac{(C_g + \beta C_d)^2 + (\beta C_g + C_d)^2}{2(C_g + \beta C_d)(\beta C_g + C_d)} - 1 + 1\right) \tag{3.178}$$

$$= \beta\left(\frac{((C_g + \beta C_d) - (\beta C_g + C_d))^2}{2(C_g + \beta C_d)(\beta C_g + C_d)} + 1\right) \tag{3.179}$$

$$= \beta\left(\frac{(C_g - C_d)^2(1 - \beta)^2}{2(C_g + \beta C_d)(\beta C_g + C_d)} + 1\right) \tag{3.180}$$

$$\tag{3.181}$$

Note that the first term in the parenthesis has two zeros at $\beta = 1$. It follows,

$$\left.\frac{\partial A_{C-H}(\beta)}{\partial \beta}\right|_{\beta=1} = 0 \Rightarrow A_{C-H} \approx \beta = \frac{g_{m1}}{g_{m2}} \tag{3.182}$$

which requires that the successive stages have $\frac{g_{m1}}{g_{m2}} = \sqrt[4]{e}$. Compared to a simple first-order amplifier,

$$GBW_{C-H,-3\text{dB}} = \sqrt{2}\frac{g_{m2}}{(C_g + \beta C_d) + (\beta C_g + C_d)}\beta \tag{3.183}$$

$$= \sqrt{2}\frac{g_{m2}}{(1+\beta)(C_g + C_d)}\beta \tag{3.184}$$

$$= \frac{\sqrt{2}\beta}{1+\beta} GBW_{0,-3\text{dB}} \tag{3.185}$$

where $GBW_{0,-3\text{dB}} = \frac{g_m}{C_g + C_d}$ is the product of gain and -3dB attenuation bandwidth. Note that for the optimal gain $GBW_{C-H,-3\text{dB}} = .62GBW_{0,-3\text{dB}}$, which shows that the Cherry-Hooper amplifier actually performs worse compared to a single stage. However, when $\beta > \sqrt{2} + 1 \approx 2.4$, the Cherry-Hooper wins over its first-order single-stage counterpart.

Since the Cherry-Hooper amplifier consists of two active components, it is also instructive to compare it to a 2-stage first-order amplifier. The Cherry-Hooper topology wins when $GBW_{C-H,-3\text{dB}} > GBW_{N=2,-3\text{dB}}$, which means that

$$\frac{\sqrt{2}\beta}{1+\beta} > A_{tot}{}^{1-\frac{1}{N}} \left.\frac{\sqrt{\alpha^{\frac{-1}{N}} - 1}}{\sqrt{\alpha^{-1} - 1}}\right|_{N=2,\alpha=\frac{1}{2},A_{tot}=\beta} \tag{3.186}$$

$$\frac{\sqrt{2}\beta}{1+\beta} > \sqrt{\beta}\sqrt{\sqrt{2} - 1} \tag{3.187}$$

This condition is satisfied as long as

$$\sqrt{2} - 1 < \beta < \sqrt{2} + 1 \tag{3.188}$$

which means that an optimally designed 2-stage first-order amplifier still outperforms the Cherry-Hooper topology, albeit only slightly. However, given the sharper out-of-band roll-off, it is better suited in a cascaded chain. Fig. 3.51 shows a comparison of the different designs.

So far, it has been shown that the advantage of the Cherry-Hooper topology is the sharper slope for out-of-band signal suppression. Shunt peaking with active inductors should be investigated as a means of improving the bandwidth of a single-stage amplifier. Using the KVL-KCL equations, the frequency response of the circuit of Fig. 3.52 can be calculated as

$$\frac{V_{out}}{V_{in}}(s) = -\frac{g_{m1}}{g_{m2}} \frac{R_f C_2 s + 1}{1 + \frac{C_1 + C_2}{g_{m2}} s + \frac{C_1 C_2 R_f}{g_{m2}} s^2} \tag{3.189}$$

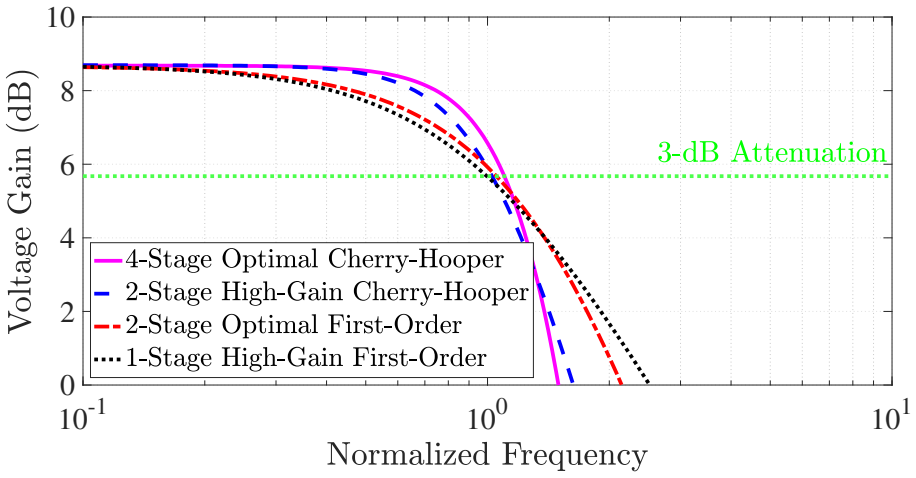


Figure 3.51: Comparison between the Cherry-Hooper topology and first-order amplifiers

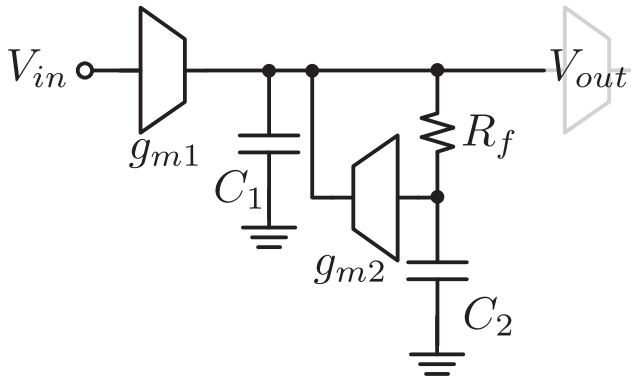


Figure 3.52: Simplified model of an amplifier with active inductor

For a maximally flat design,

$$\left.\frac{\partial \left|\frac{V_{out}}{V_{in}}(\omega)\right|}{\partial \omega}\right|_{\omega=0} = 0 \tag{3.190}$$

$$\left.\frac{\partial^2 \left|\frac{V_{out}}{V_{in}}(\omega)\right|}{\partial \omega^2}\right|_{\omega=0} = 0 \tag{3.191}$$

$$\left.\frac{\partial^3 \left|\frac{V_{out}}{V_{in}}(\omega)\right|}{\partial \omega^3}\right|_{\omega=0} = 0 \tag{3.192}$$

The first and third derivatives are always satisfied. The second derivative is satisfied when

$$(R_f C_2)^2 = \frac{(C_1 + C_2)^2}{{g_{m2}}^2} - \frac{2C_1 C_2 R_f}{g_{m2}} \tag{3.193}$$

and

$$g_{m2} R_f = \frac{\sqrt{{C_1}^2 + (C_1 + C_2)^2} - C1}{C_2} \tag{3.194}$$

Note that the DC gain is given by $\frac{g_{m1}}{g_{m2}} = \beta$. So let us assume that the chain has a per-stage fan out of f,

$$C_1 = \beta C_d + C_d + f C_g \tag{3.195}$$
$$C_2 = C_g \tag{3.196}$$

For most practical cases, $C_2 \ll C_1$ and

$$g_{m2} R_f \approx \frac{\sqrt{2}}{2} + (\sqrt{2} - 1)\frac{C_1}{C_2} \tag{3.197}$$

The definition of a natural frequency is more complicated here since this system has one zero and two poles. Instead, we use the natural frequency of the similar Butterworth response, which has the same 4-th derivative

$$\omega_n : \left.\frac{\partial^4 \left|\frac{V_{out}}{V_{in}}(\omega)\right|^2}{\partial \omega^4}\right|_{\omega=0} = \left.\frac{\partial^4 \left(\frac{1}{1+\left(\frac{\omega}{\omega_n}\right)^4}\right)}{\partial \omega^4}\right|_{\omega=0} \tag{3.198}$$

Substitute the required R_f into the gain equation and take the derivatives,

$$\frac{12{C_1}^2\left(-3{C_1}^2 + \left(2\sqrt{2{C_1}^2 + 2C_1C_2 + {C_2}^2} - 2C_2\right)C_1 - {C_2}^2\right)}{{g_{m2}}^4} = \frac{-24}{{\omega_n}^4} \tag{3.199}$$

Assuming $C_2 \ll C_1$, the natural frequency can be approximated as

$$\omega_n \approx \frac{g_{m2}}{C_1}\frac{\sqrt{2}}{\sqrt[4]{6 - 4\sqrt{2}}}\left(1 - \frac{1}{4 - 2\sqrt{2}}\frac{C_2}{C_1}\right) \tag{3.200}$$

and the -3dB gain-bandwidth product is

$$GBW_{ActiveInd} \approx \frac{g_{m1}}{C_1}\frac{\sqrt{2}}{\sqrt[4]{6 - 4\sqrt{2}}}\left(1 - \frac{1}{4 - 2\sqrt{2}}\frac{C_2}{C_1}\right) \tag{3.201}$$
$$\approx GBW_0\frac{\sqrt{2}}{\sqrt[4]{6 - 4\sqrt{2}}}\left(1 - \frac{1}{4 - 2\sqrt{2}}\frac{C_2}{C_1}\right) \tag{3.202}$$

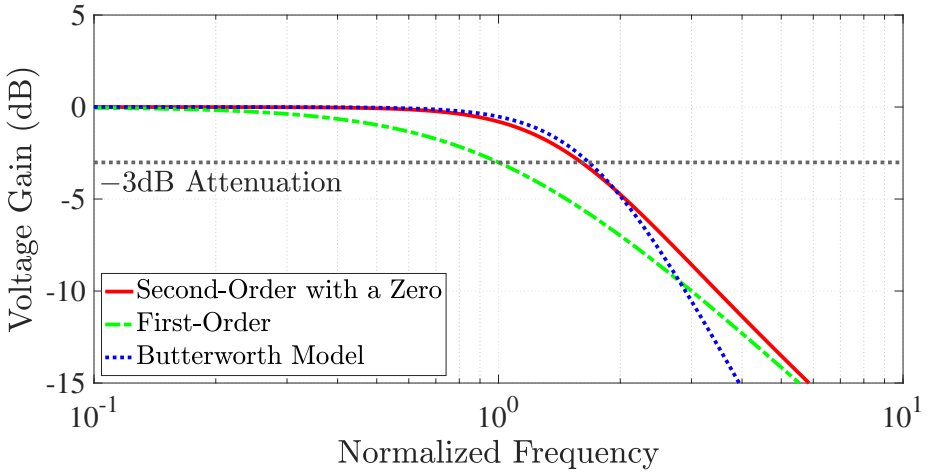


Figure 3.53: Comparison of the voltage gain in an amplifier with active inductor with its first-order and Butterworth counterparts

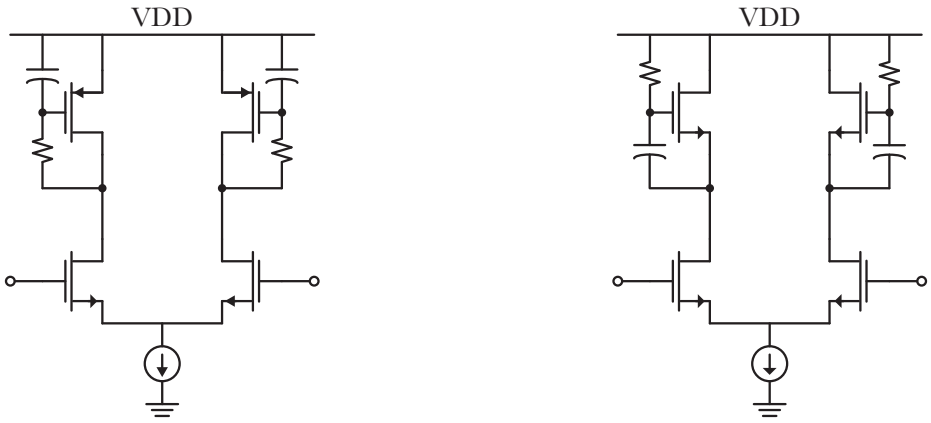


Figure 3.54: PMOS and NMOS implementation of the active inductor

Note that the gain-bandwidth product increases by about 85% compared to a simple first-order stage. Fig. 3.53 shows that the Butterworth model used here agrees well with the actual transfer function.

Having demonstrated the effectiveness of an active inductor to increase the bandwidth of each stage, PMOS and NMOS active inductors are investigated, as in Fig. 3.54. The problem with these topologies is that the current density of the active load must be higher than that of the differential pair to achieve gains greater than 1. This means that for a fixed GBW_0 of the differential pairs, a higher parasitic C_2 can be expected, which lowers the $GBW_{ActiveInd}$. To obtain the near-optimal current density for all devices, the topology of Fig. 3.55 is chosen. In this design, the active devices have the near-optimal current density for maximum speed. Moreover, both bandwidth and gain are controllable by triode devices (purple and green transistors, respectively). At the maximum gain setting

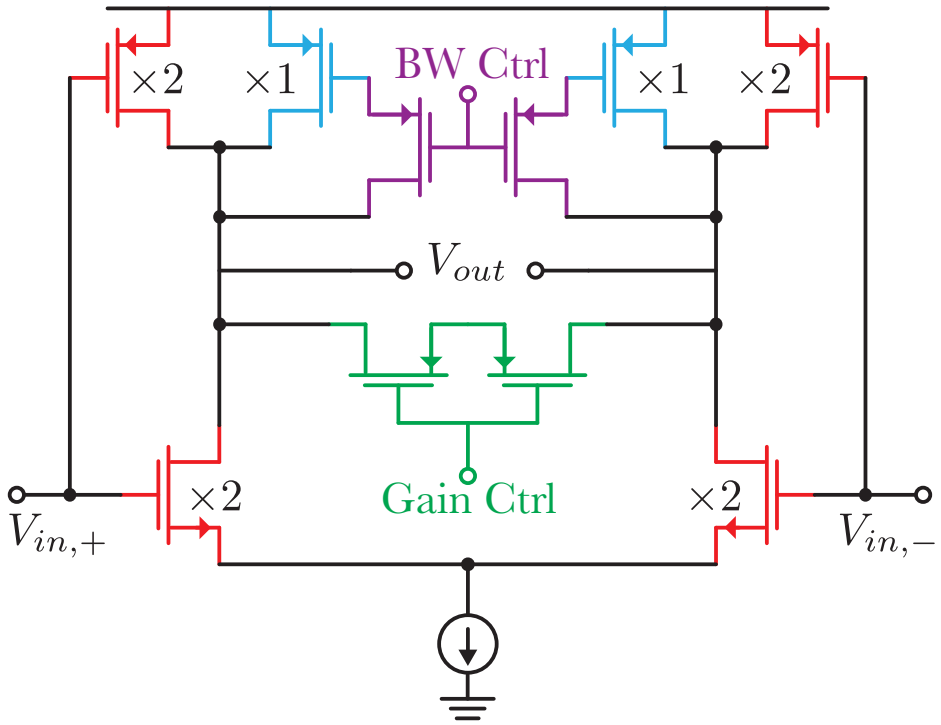


Figure 3.55: Final implementation of the amplifier with active inductor

$$G = (g_{m1} + g_{m2})\left(g_{o1} + g_{o2} + g_{o3} + g_{m3}\right)^{-1} \tag{3.203}$$

With the same mobility and intrinsic gain for PMOS and NMOS devices, the DC gain can be calculated as

$$I_{m2} = \frac{2}{3} I_{m1} \Rightarrow g_{m2} \approx \sqrt{\frac{1}{1} \times \frac{2}{3}} g_{m1}, g_{o2} \approx \sqrt{\frac{1}{1} \times \frac{2}{3}} g_{o1} \tag{3.204}$$

$$I_{m3} = \frac{1}{3} I_{m1} \Rightarrow g_{m3} \approx \sqrt{\frac{1}{2} \times \frac{1}{3}} g_{m1}, g_{o3} \approx \sqrt{\frac{1}{2} \times \frac{1}{3}} g_{o1} \tag{3.205}$$

Where the transconductance of the device is $gm = \sqrt{2\mu C_{ox} \frac{W}{L} I_{DC}}$. Hence,

$$G = g_{m1}\left(1 + \sqrt{\frac{2}{3}}\right) r_{o1} \left(1 + \sqrt{\frac{2}{3}} + \sqrt{\frac{1}{2} \times \frac{1}{3}} + \sqrt{\frac{1}{2} \times \frac{1}{3}} g_{m1} r_{o1}\right)^{-1} \tag{3.206}$$

which corresponds to about 2.8 for an intrinsic gain of 10. This is very close to the optimal gain in a power-efficient cascaded chain for a total gain of 30dB and a fanout factor of 10. Fig. 3.56 shows the simulated performance of the cascaded chain. Table. 3.2 compares this work with previous work. Note that both DC gain and fan-out factor should be considered for a fair comparison between different results.

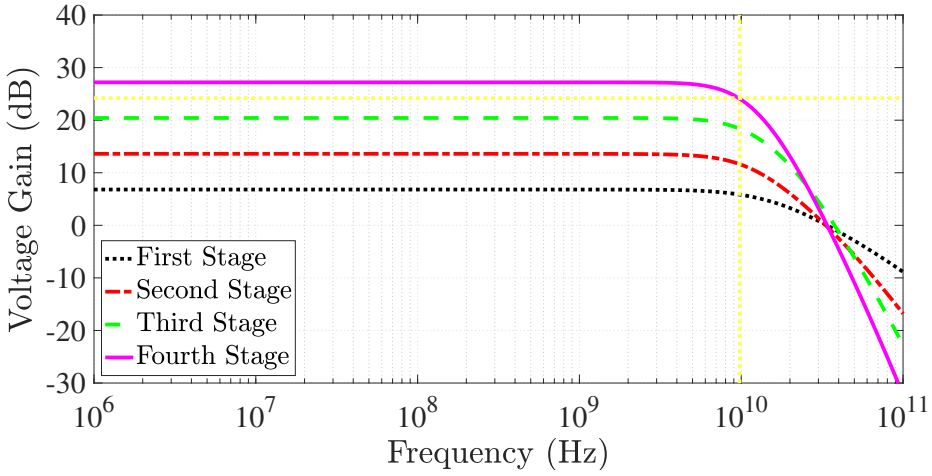


Figure 3.56: Performance of cascaded active inductor stages

	CMOS Tech.	Bandwidth	Gain	DC Power	Fan-out
[35]	28nm	19.2GHz	28.3dB	10.3mW	1
This	28nm	9.64GHz	27.2dB	10.6mW	7

Table 3.2: Comparison of the baseband amplifier with earlier work

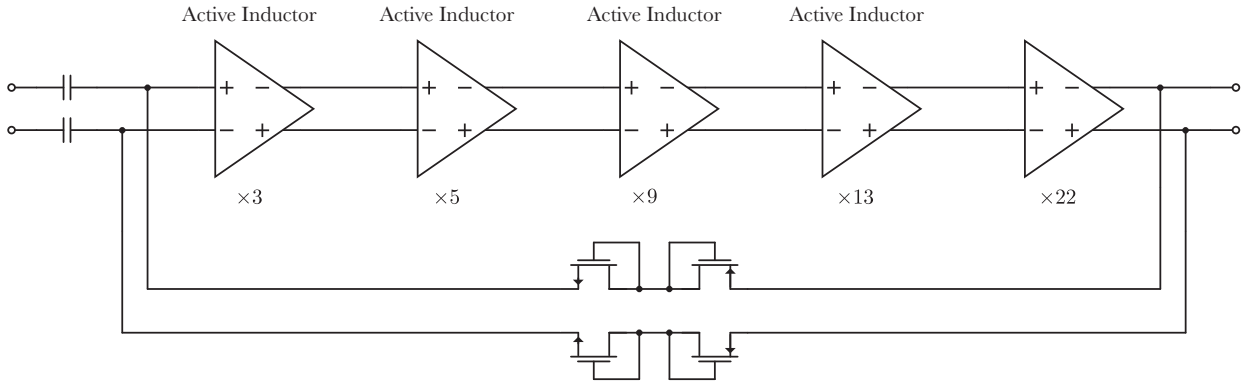


Figure 3.57: Baseband chain

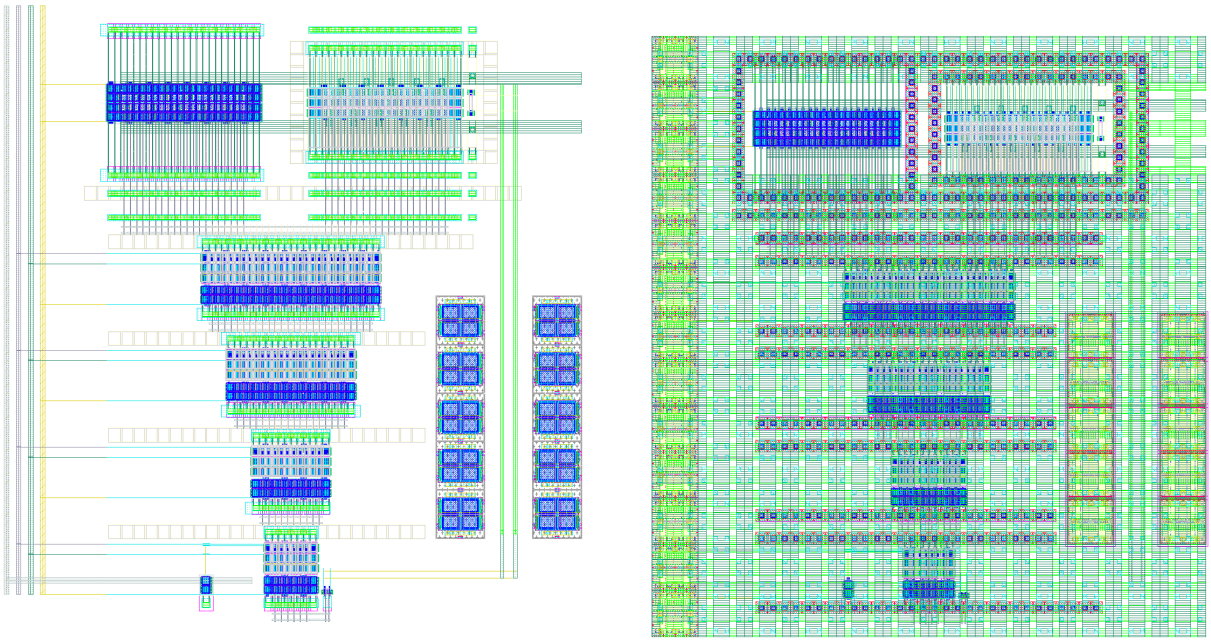
The disadvantage of this design is that the output swing is limited. Therefore, an additional stage is inserted that does not contain an active inductor, as in Fig. 3.57. The differential feedback provides the DC bias for the chain. Remember that common-mode feedback is necessary for an amplifier with a differential input and output. However, for this

amplifier, the common-mode gain can be approximated as

$$G_{cm} \approx g_{m2}\left(g_{o2}+g_{o3}+g_{m3}\right)^{-1} \tag{3.207}$$

$$\approx \sqrt{\frac{2}{3}}g_{m1}r_{o1}\left(\sqrt{\frac{2}{3}}+\sqrt{\frac{1}{2}\times\frac{1}{3}}+\sqrt{\frac{1}{2}\times\frac{1}{3}}g_{m1}r_{o1}\right)^{-1} \tag{3.208}$$

which is approximately 1.8 for an intrinsic gain of 10. This common-mode gain is achieved by using an odd number of stages to ensure that the feedback polarity is negative once the loop is closed. To avoid loop compensation, the feedback resistor is implemented using triode devices. This results in the dominant pole of the feedback loop being at the input of the chain. Fig. 3.58 shows the layout of the baseband amplifier. Note that to avoid latch-up



(a) Simplified Layout (b) Actual Layout

Figure 3.58: The layout of the baseband amplifier

and ESD failures, the last stage is implemented with individual guard rings for each set of PMOS and NMOS.

To cope with the ESD and pad capacitance, series inductors are used to form an artificial transmission line, as shown in Fig. 3.59. Due to the limited area available for the tape out and the congestion of the phased array units, the final performance is suboptimal. Fig. 3.60 shows the performance of the entire baseband chain, including the pads and ESD units. The entire baseband chain consumes 20mW.

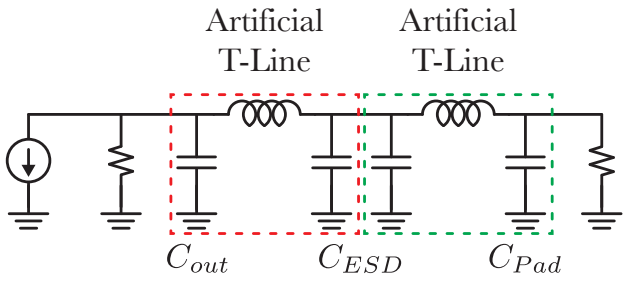


Figure 3.59: Using an artificial T-line to increase the bandwidth

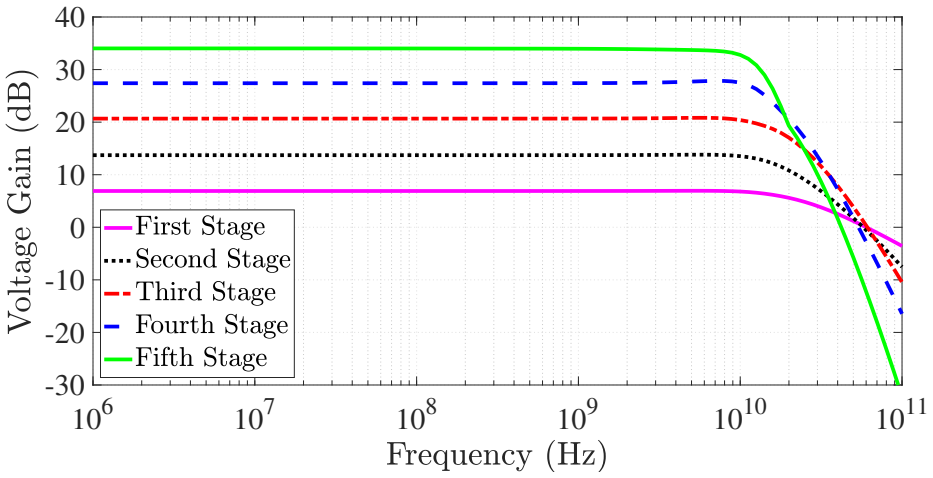
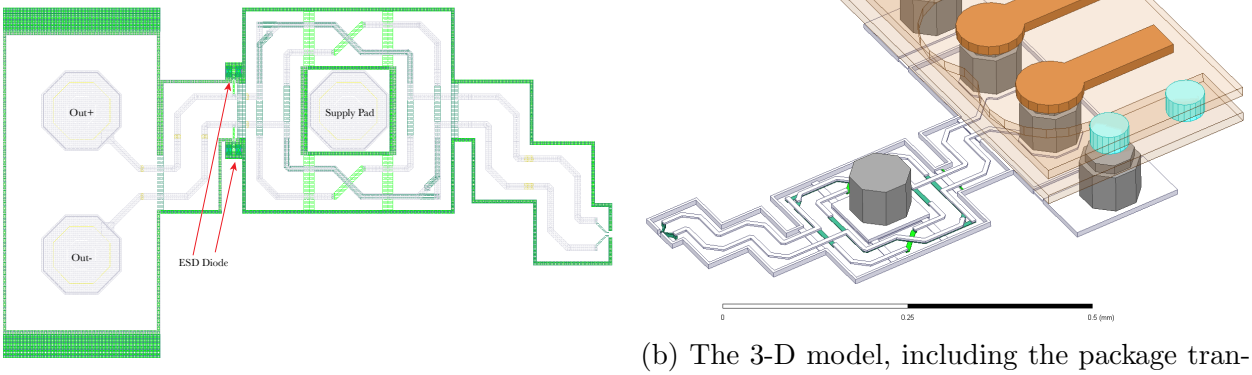


Figure 3.60: Baseband Chain Performance

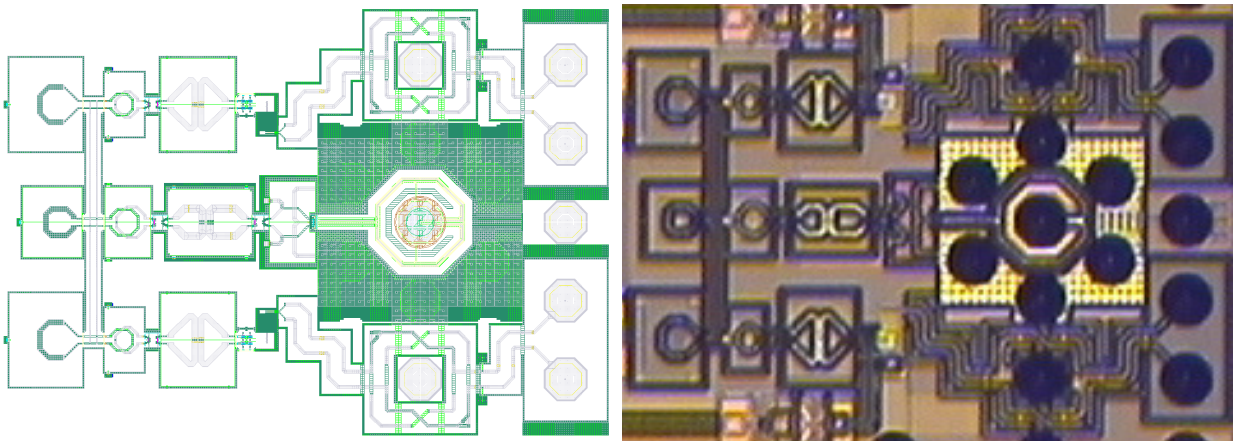


(a) Layout View

(b) The 3-D model, including the package transition

Figure 3.61: The layout of the baseband amplifier

3.9 Full Receiver Performance



(a) Receiver layout (b) Die photo

Figure 3.62: 140GHz receiver taped out in 28nm CMOS technology.

This receiver is implemented in a 28nm Bulk CMOS process. The die photo and layout of the chip are shown in Fig. 3.62. The receiver consumes 60mW power, details are in Fig. 3.63.

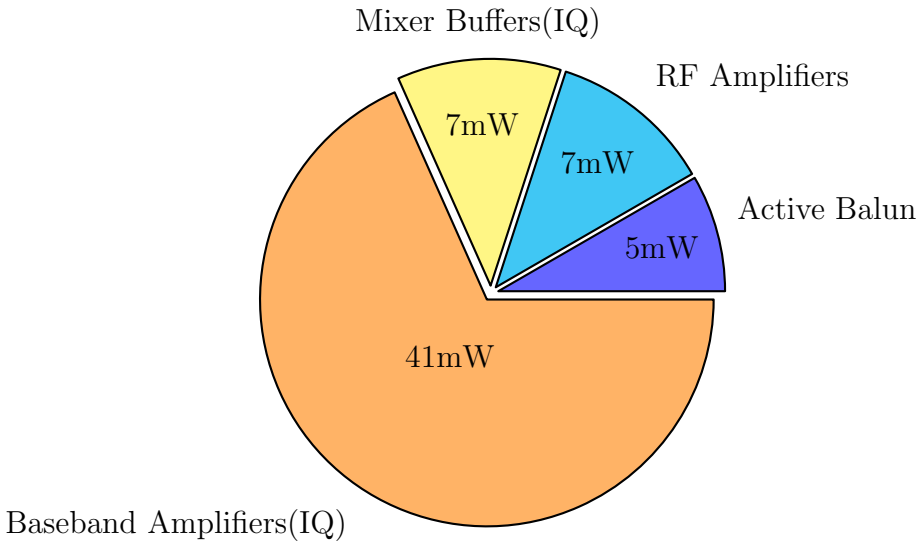


Figure 3.63: Power consumption of the receiver

Fig. 3.64 shows the performance of the receiver chain. While the 3dB bandwidth of the output is 11GHz, a bandwidth of 18GHz with a noise figure of 3dB is achievable when equalization is applied. Details of the performance can be found in Table. 3.3, where this

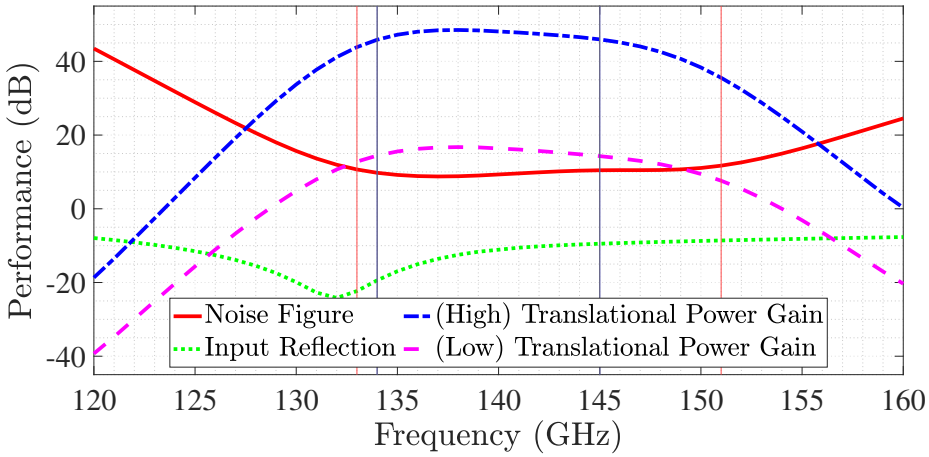


Figure 3.64: Performance of the receiver chain

work is compared with other published work. SOI processes have better performance due to superior devices and RF-optimized back-end metallization.

	This Work	[37]	[19]	[38]	[39]	[40]	[20]
Technology	28nm CMOS	65nm CMOS	45nm CMOS SOI	45nm CMOS SOI	45nm CMOS SOI	22nm CMOS SOI	28nm CMOS
Carrier Frequency (GHz)	140	140	147	144	140	135	113
RF Bandwidth (GHz)	11	20	16	14	12	20	10
RX Gain (dB)	48	43	27.5	26.5	18	27	43.8
RX NF (dB)	10	11	6.4	6.4	5.5	8.5	11.2
Power Consumption (W)	0.060	NA	0.145	0.133	0.125	0.198	0.500

Table 3.3: Comparison of the receiver with the state-of-the-art

Fig. 3.65 shows the gain of the chain as a function of the input power. Note that the linearity of the circuit is mainly limited by the output swing of the baseband amplifier in the high and low gain modes.

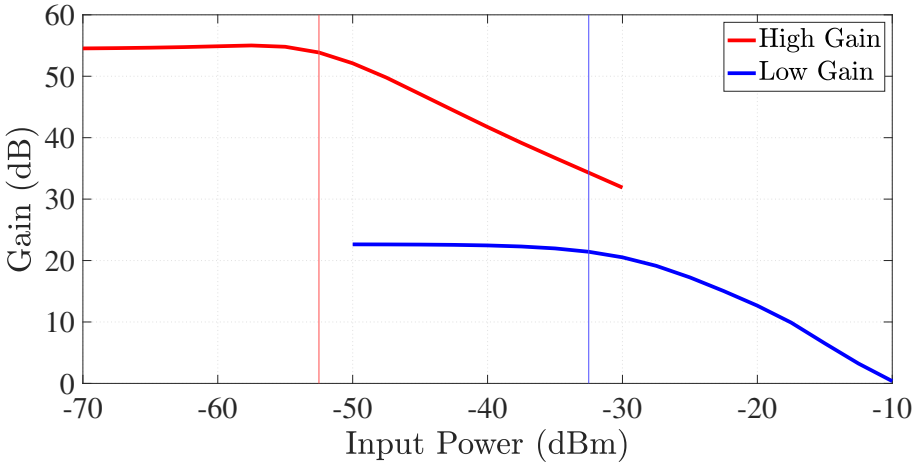


Figure 3.65: Translation gain vs. input power

Chapter 4

Chip-to-Package Transition

4.1 Packaging Challenges at High Frequencies

The transition of the signal from the chip to the printed circuit board (PCB) becomes increasingly difficult as the carrier frequency increases. Several factors play a role in this:

- Although wire-bond is still the primary packaging solution for the frequency range above 100GHz, it is not reliable for massive array deployment. Since PCB fabrication capabilities dictate wire-bond length [41], the parasitic inductance of wire-bonds reduces the achievable bandwidth even with tuning techniques [42]. On the other hand, the horizontal alignment and vertical dimensions of flip-chip technology can be controlled with an accuracy of ten microns or less [43].

- Most PCBs have limited resolution in trace spacing and trace width. As a rule of thumb, for a transmission line, the return path should be closer than $\frac{\lambda_{min}}{10}$ to the signal path, where λ_{min} is the wavelength at the maximum operating frequency. With a trace spacing of 6mil $\approx 150\mu$m on an FR-4 dielectric (with a relative dielectric constant of 4), transmission lines are limited to a maximum frequency of 100GHz.

- The diameter of the bumps or studs used for the transition determines the minimum distance between signal and return current. AuSn micropumps with a diameter of 10μm, for example, have shown return loss of better than 10dB up to 250GHz [44]. Unfortunately, these small bumps are costly and require much higher accuracy in PCB fabrication and chip assembly. As the spacing and diameter of the balls increase, unintended resonant modes can create notches near the band of interest.

- If the transition is not properly shielded, it can become a radiating element. For example, at a pitch of 150μm, two pads become a radiating dipole at

$$f_{rad} = \frac{v_0}{2\sqrt{\epsilon_{si}}150\mu\text{m}} = 290\text{GHz} \tag{4.1}$$

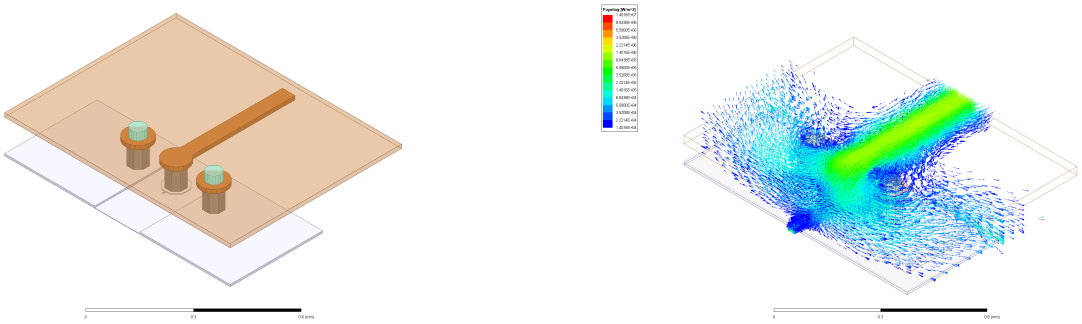
- Metal planes on the PCB adjacent to the metal planes of the chip support the parallel-plate propagation mode. Once the new wave is excited, it is reflected from nearby bumps, or partially radiated and partially reflected from the chip boundaries. The reflected wave changes the effective impedance at the excitation point. In addition, if the reflected wave adds destructively with the original wave, notch behavior occurs in the transfer characteristic.

The suitability of flip-chip packages for millimeter-wave applications is well explained in [43]. Furthermore, various non-idealities occurring in flip-chip packages are described in [45]:

- Detuning: the presence of a semiconductor dielectric on the PCB changes the effective dielectric constant on the transmission lines [45]. Therefore, it is necessary to keep high-frequency I/Os at the periphery of the chip and minimize the distance between the signal pad and the chip edge. In addition, it is often essential to use an underfill between the chip and the board to increase the mechanical reliability of the assembled chip. While the volume of the added underfill is controlled, its exact shape and extension beyond the chip edge are unknown, making it difficult to model its loss and detuning effect properly.
- Excitation of parasitic modes: Considering only the semiconductor and its metal plane, this structure supports the propagation of TE and TM waves, commonly referred to as surface modes. While a lower thickness of the semiconductor shifts the cut-off frequency of these parasitic modes to a higher frequency range, the mechanical strength of the chip is reduced, leading to a higher susceptibility to mechanical stress. Unfortunately, these higher-order modes are always excited at the boundary of the chip where the signal transition occurs. When resonance occurs, the transition can have a very high loss.
- Reflections and insertion loss at the transition site.

4.2 Transition Structures

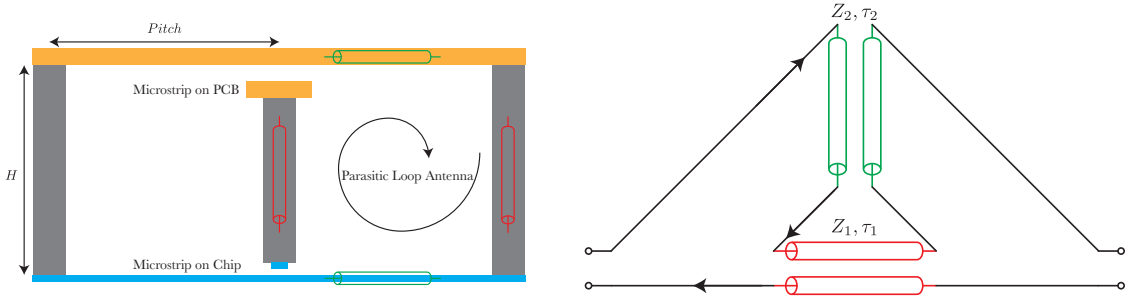
The transition structure of most works is still a simple ground-signal-ground (GSG) structure. Therefore, the only method to improve the transition performance is to use smaller bumps. Here, different structures are analyzed to improve the performance with a fixed bump diameter of 75μm.



(a) Structure (b) Poynting Vector at 400GHz

Figure 4.1: Conventional microstrip GSG pads

Let us start the analysis with a simple structure shown in Fig. 4.1a. The cross-section of this transition is shown in Fig. 4.2a. Intuitively, we can see that the time for the signal current



(a) Symbolic model (b) Schematic model

Figure 4.2: Modeling the microstrip transition with transmission lines

and the return current to travel from the PCB to the chip are not the same. The extra length of metal and the corresponding delays can be modeled with transmission lines. In this model, the bumps that carry the current in the vertical direction are intentional transmission lines (red in Fig. 4.2a). In contrast, the horizontal paths that the return currents must follow on the PCB and chip are parasitic transmission lines (green in Fg. 4.2a). In the simple transmission line model of Fig. 4.2b, the input current into the intended transmission line (labeled 1) must equal the input current into the parasitic line (labeled 2). Therefore,

$$I_{in} = \frac{v_{2f} - v_{2r}}{Z_2} = \frac{v_{2f}e^{-j\theta_2} - v_{2r}e^{j\theta_2}}{Z_2} \tag{4.2}$$

Where v_{xf} and v_{xr} are the voltage of the propagating waves in the forward and reverse directions in each transmission line. To satisfy this equation,

$$e^{j\theta_2} = -\frac{v_{2f}}{v_{2r}} \tag{4.3}$$

Interestingly, the standing wave ratio on the second transmission line does not depend on the load impedance. Moreover, at the frequency where $\theta_2 = \pi$,

$$\begin{aligned} I_{in} &= \frac{v_{2f} - v_{2r}}{Z_2} \\ &= \frac{v_{2f} + v_{2f}e^{-j\theta_2}}{Z_2} \\ &= 0 \end{aligned} \tag{4.4}$$

which suggests that at the frequency of

$$f_{notch} = \frac{1}{2\tau_2} \tag{4.5}$$

or an odd integer multiple of that, a notch in the transmission characteristic is expected. In other words, the timing mismatch between current and reverse current results in deep notches in the transition. Note that under the assumption of loss-less transmission lines, even near the notch frequency, G_{max} remains high because any reactive energy can be tuned out with ideal components, at least in theory. However, the tuning comes at the cost of extremely low bandwidth and high insertion loss due to the matching elements. The other transmission line may also exhibit similar notch behavior; however, for most practical transitions $\tau_1 < \tau_2$. Note that this deep notch is easily seen when the length of the horizontal line is much greater than that of the vertical line, which is usually the case when small bumps are used on low manufacturing resolution PCBs.

Since the green transmission line is a 2-D parallel-plate transmission line, the signal escapes by coupling to the parallel-plate propagation mode at the metal-dielectric-metal stack in the transition region. Fig. 4.1b shows the direction of the Poynting vector. If we assume an optimal situation, the length of the two transmission lines should be similar. Moreover, a parasitic loop antenna is excited at the transition in this situation. The loop antenna is in resonance when the circumferential length of the loop is equal to the wavelength, assuming a short circuit on the chip. In terms of delays in the transmission line model

$$f_{rad} = \frac{1}{2(\tau_1 + \tau_2)} \tag{4.6}$$

The incoming signal near the radiation frequency is dissipated by coupling with parasitic surface wave modes and parallel plate modes. This factor is clearly seen when G_{max} is considered. Fig. 4.3 shows G_{max} for different distances as a function of frequency. Here, a bump height of 75μm is considered, which is the minimum bump height offered by the technology used. It can be observed that as the bump height increases, the first notch moves closer to the origin. In the simulation structure, the total distance between two footprints (H in Fig. 4.2a) is 125μm. With a dielectric constant of 3.1 for the underfill material, Eq. 4.6 estimates the first notch to be

$$f_{notch} \approx \frac{1}{2} \frac{\frac{3\times10^8 \mathrm{ms}^{-1}}{\sqrt{3.1}}}{125\mu\mathrm{m} + Pitch} \tag{4.7}$$

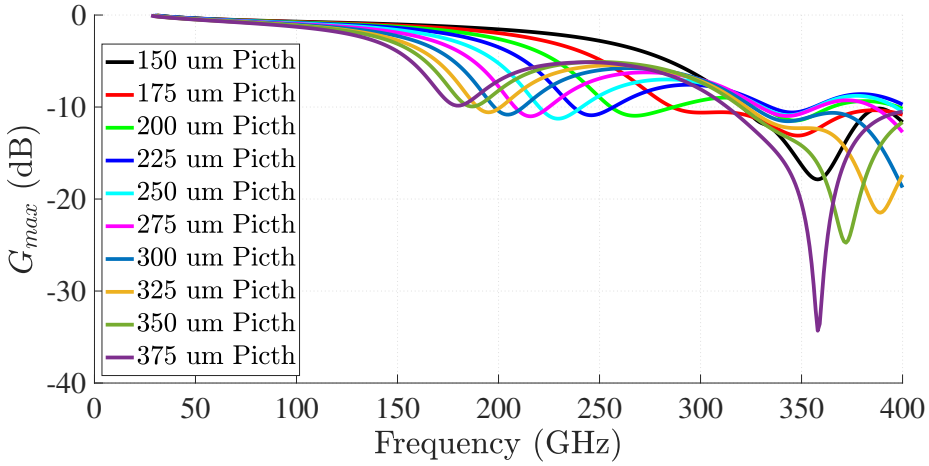


Figure 4.3: G_{max} versus frequency for different distances

where *Pitch* in Fig. 4.2a. As you can see in Fig. 4.4, the radiation frequency of the loop antenna agrees well with the simulation results.

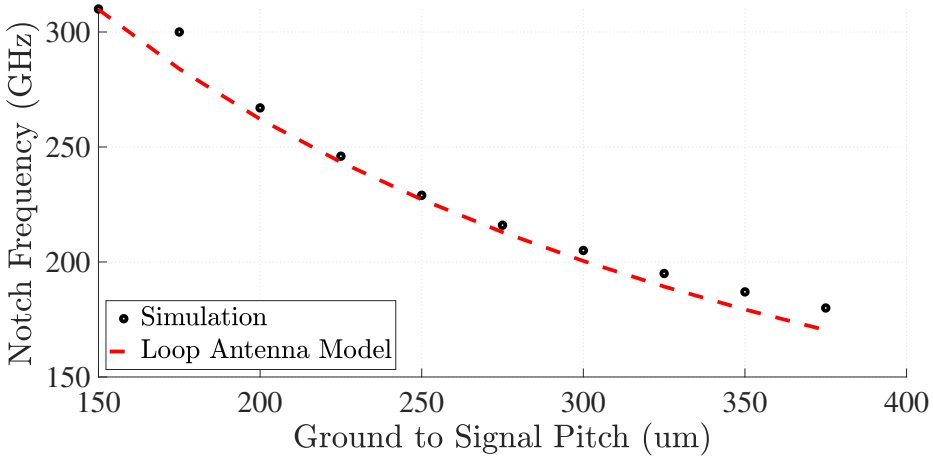
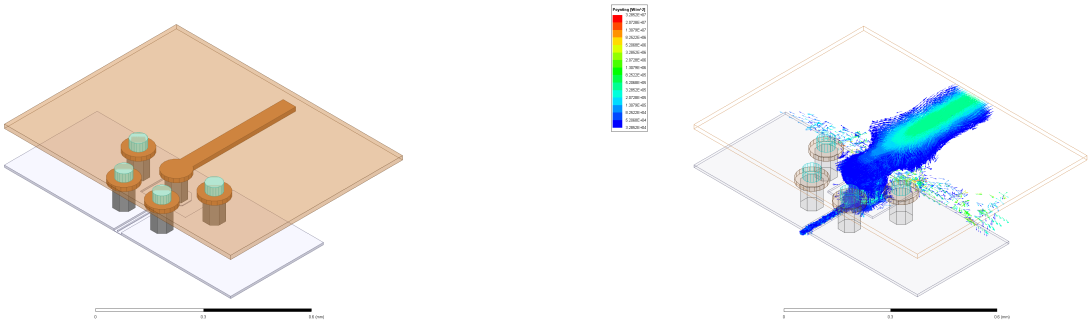


Figure 4.4: Notch frequency of G_{max} in the simulation versus the loop antenna model

Let us now discuss some other structures.

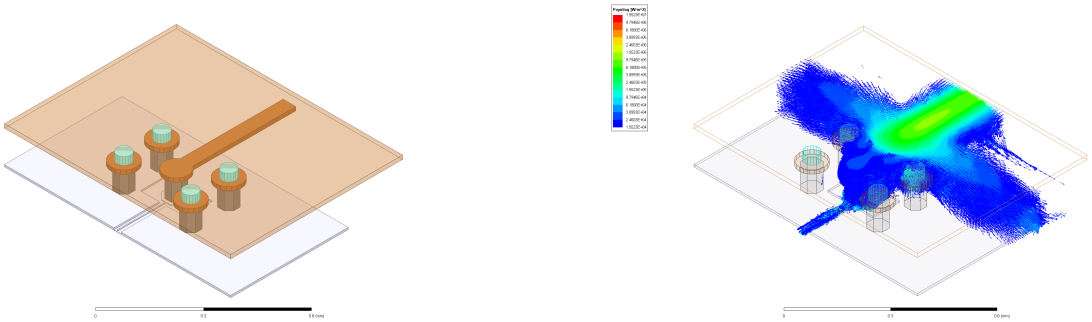
- By adding additional ground bumps as in Fig. 4.5a, one can partially reflect surface waves, which should reduce the transition loss. However, the reflected wave will still reach the other side of the chip and will be dissipated either by radiation or excitation of surface waves across the chip boundary. As you can see in Fig. 4.10b, this method is quite effective in reducing the transition loss at the previous radiation frequency and shifting the notch to a higher frequency.



(a) Structure (b) Poynting Vector at 400GHz

Figure 4.5: Microstrip with front shield

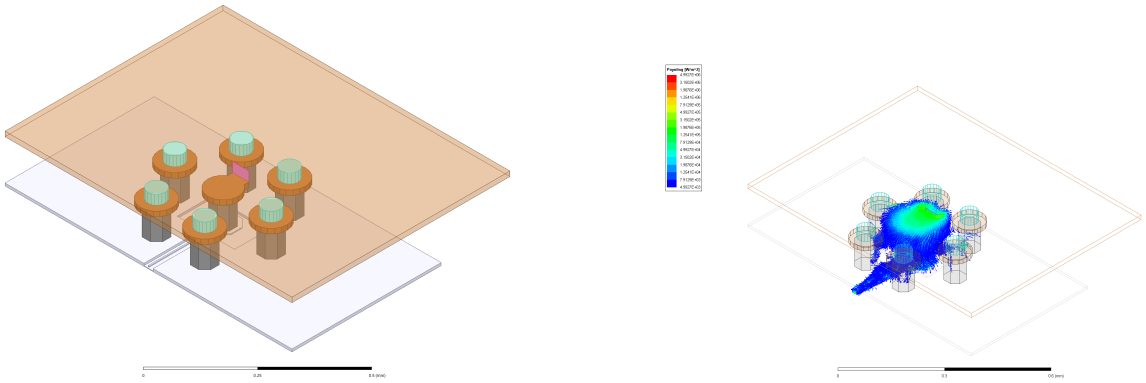
- By adding two sets of ground bumps with positive and negative offsets in a rectangular shape, one can make the transition as shown in Fig. 4.5a. This is a much better approach in the lower frequency range because it can effectively reject forward and backward surface waves. However, as the distance between two ground bumps increases, higher leakage is expected, as shown in Fig. 4.10b. Moreover, as the length of the PCB microstrip line increases over the chip region, this structure suffers from a higher degree of detuning and coupling with the silicon substrate. This indicates that the least leakage is expected when a full bump cage is formed with minimal spacing.



(a) Structure (b) Poynting Vector at 400GHz

Figure 4.6: Microstrip with rectangular shield

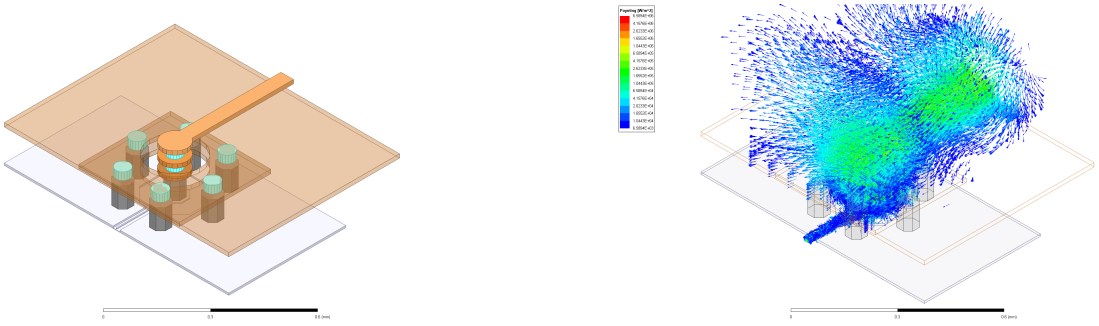
- As mentioned earlier, the best performance is expected when the smallest spacing between all bumps is used. To achieve this goal, the ground bumps must be on a hexagon around the signal, as shown in Fig. 4.7a. The simulation results shown in Fig. 4.10b indicate that this structure achieves the best performance in terms of transition loss and notch frequency. Unfortunately, depending on the capabilities of the PCB manufacturer, this design may be impractical since the microstrip signal must be squeezed out of two ground bumps and their associated pads.



(a) Structure (b) Poynting Vector at 400GHz

Figure 4.7: Microstrip with full shield

- If the previous structure with a full shield was not practical, a reverse microstrip could be used, as in Fig. 4.8a. In this case, the metal layer of the microstrip is on the top-most layer, while the signal metal is buried underneath. This strategy allows us to place the feedline in the middle of the chip. This additional degree of freedom will enable us to use the periphery of the chip for other purposes. However, it requires interruptions on the transmission line's ground plane to accommodate more I/Os. To avoid this interruption, the ground plane of the inverted microstrip can be implemented on the second top layer while the signal is on the third layer. This transition topology minimizes leakage at the chip interface. However, signal loss occurs at the inner via. The simulation results (Fig. 4.10b) show that this structure has higher losses compared to the other topologies. Moreover, it requires a large keep-out region above the signal line to reduce the parasitic coupling, making it less attractive.

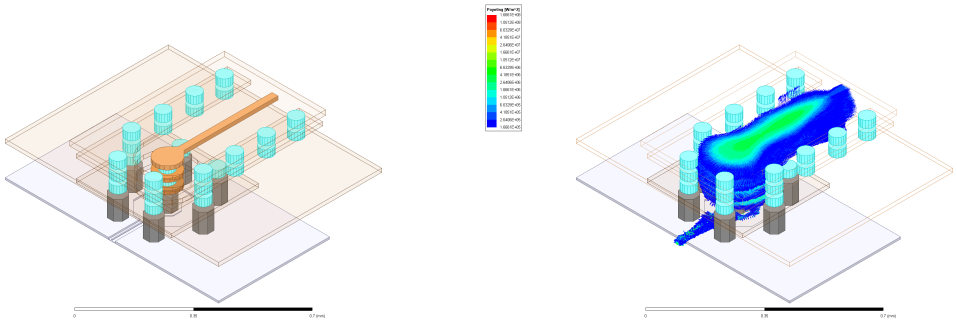


(a) Structure (b) Poynting Vector at 400GHz

Figure 4.8: Reverse microstrip with full shield

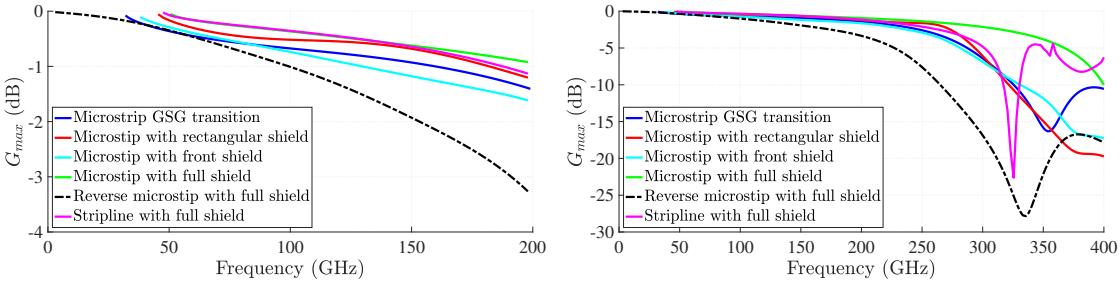
- To solve the previous problem, the microstrip line can be replaced by a stripline, as shown in Fig. 4.9a. The simulation results (Fig. 4.10b) show that this structure has

superior performance compared to other practical options up to 220GHz. After that, the transition's loss increases with increasing frequency, and at 325GHz, there is a notch in the transmission characteristic.



(a) Structure (b) Poynting Vector at 400GHz

Figure 4.9: Stripline with full shield

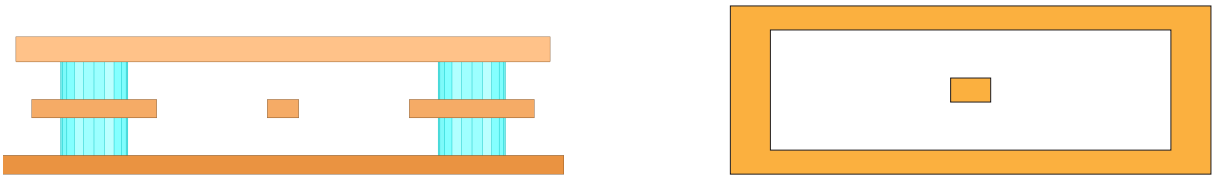


(a) Sub 200GHz (b) DC to 400GHz

Figure 4.10: G_{max} of the different transition scenarios

4.3 Limitation of the Stripline Structure

So far, the stripline design of Fig. 4.9a is the most promising solution for high frequencies. Another advantage of this topology is that the millimeter-wave signal is completely shielded from the environment. This means that the performance is less susceptible to variations in the shape of the underfill or the expansion of the silicon. Therefore, it is desirable to explore this structure and investigate its possible limitations. First, the PCB stripline itself should be investigated. The cross section of the stripline is shown in Fig. 4.11. The first propagation mode of this structure (Fig. 4.12a) is the intended TEM mode, which has no cut-off frequency. However, as the frequency increases, the metal cage around the line forms an effective waveguide, commonly called a substrate-integrated waveguide [46, 47]. Note



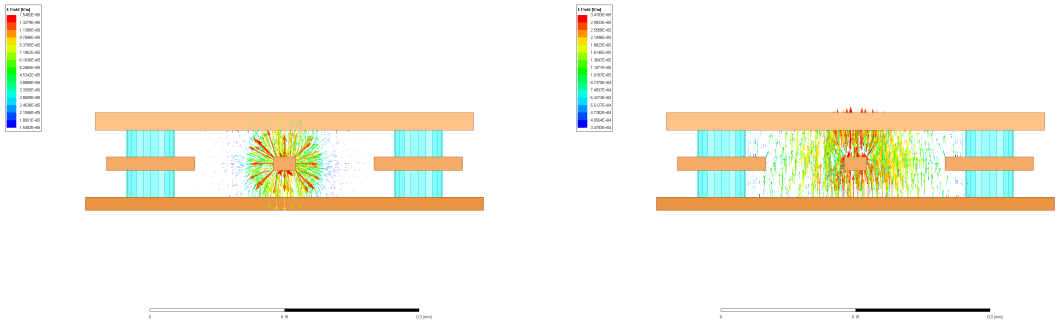
(a) Actual cross section (b) Simplified cross section

Figure 4.11: Cross section of a stripline on PCB

that the discrete nature of microvias allows only TE propagation modes in the waveguide. The effective width of the waveguide can be approximated by [48]

$$W_{eff} = W - \frac{D^2}{0.95P} \tag{4.8}$$

where W is the center-to-center spacing of the microvias on two sides, D is the diameter of the vias, and P is the spacing of the vias on the same side. The E-field of the first TE mode of this effective waveguide is shown in Fig. 4.12b. Intuitively, above the cut-off frequency of the TE wave, the upper and lower ground planes may propagate different signals, indicating that the ground planes above the cut-off frequency are undefined. Fig. 4.13 shows that a

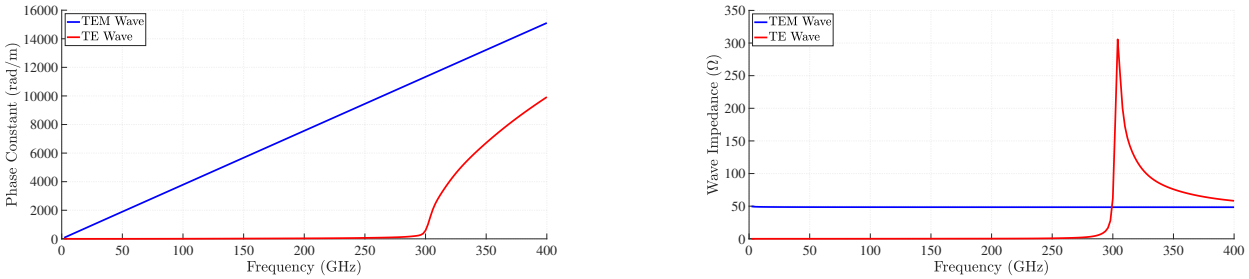


(a) TEM wave (b) TE wave

Figure 4.12: Two propagation modes of the stripline cross section

cut-off frequency of 300GHz is expected for the TE wave. This means that while an ideal straight stripline will perform smoothly in a simulation platform, any other structure may exhibit unpredictable performance if the exact length of the transmission lines is not known at the design stage. Therefore, the designer should ensure that the cut-off frequency of the TE wave is well above the highest frequency range of interest.

Considering Fig. 4.10b, the notch frequency of the stripline structure is above 300GHz. It is still important to understand the formation mechanism of this notch since process variations can change its frequency. When it is shifted to the lower frequency range, the insertion loss of the transition can increase rapidly. Let us first understand how a notch in G_{max} occurs and why it is different from a notch in transmission (S_{21}). Consider a



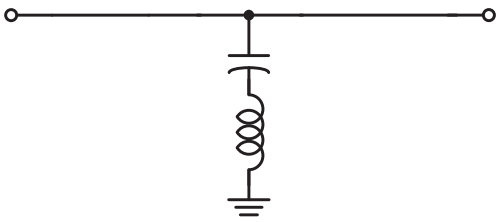
(a) Phase Constant (b) Wave Impedance

Figure 4.13: Characteristics of TEM and TE waves in striplines

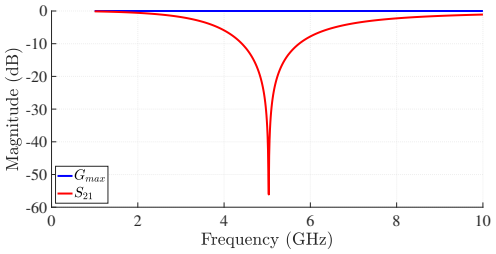
simple circuit shown in Fig. 4.14a. Note that at the resonant frequency of the tank $S_{21} = 0$. However, there is always an ideal matching network near the resonant frequency that cancels the effect of the tank. This means that $G_{max} = 0$dB over the entire frequency range (Fig. 4.14b). Intuitively, such a matching network must translate the input impedance of each port to a much lower impedance so that the equivalent parallel impedance of the tank looks much smaller than the port impedance. Translating the port impedance to a lower impedance requires passive current gain. The impedance translation ratio increases as the frequency gets closer to the notch frequency, requiring more current gain. Now considering the series loss in accessing the tank, as in Fig. 4.14c, a higher current gain increases the power loss. Therefore, as the frequency approaches the resonant frequency of the tank, the insertion loss approaches infinity, leading to a notch in G_{max} (Fig. 4.14d). The same considerations can be applied to a series tank, as in Fig. 4.14. The critical point here is that in the presence of any resonant structure, the series and parallel losses of the access lines may force $G_{max} = 0$.

To avoid such a notch, one must intentionally change the resonant frequency or ensure that the resonant structure is not excited. An eigenmode solver of Ansys HFSS was used to study the resonant modes, and the structure was modified to remove the access transmission lines. Among the numerous resonant modes, one of the modes corresponds to the cavity where the signal goes down through microvias in the shielded cage. The stripline is connected to the body of the cage, and based on the field vectors, the resonant mode of the cavity couples to the TE mode of the parasitic stripline waveguide. Therefore, depending on the reflection phase of the coupled wave, the resonant frequency of the loaded structure changes slightly. The actual reflection phase is unknown because this parasitic mode is not necessarily terminated with an actual load. Therefore, the waveguide is short-circuited at the end of the stripline, and several different lengths of the stripline are simulated (Fig. 4.15). Note that the phase constant of the TE mode approaches 0 near the cut-off frequency of the waveguide. Once the cavity's resonant frequency is shifted down towards the cut-off frequency of the waveguide, the phase shift of the reflected wave becomes independent of the length.

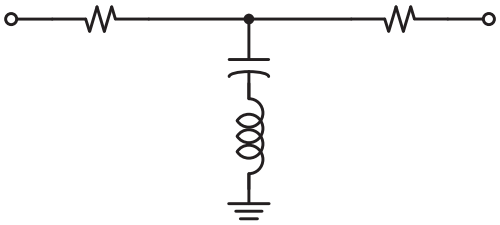
Although the notch frequency of G_{max} may shift to lower frequency bands, it will not



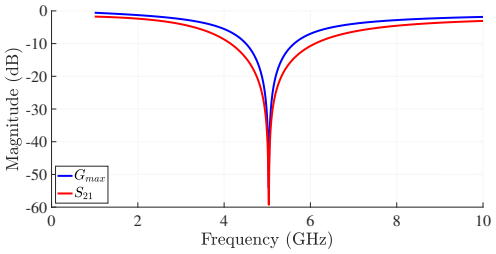
(a) Parallel LC



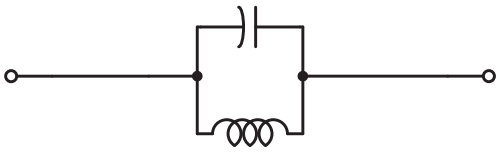
(b) Parallel LC performance



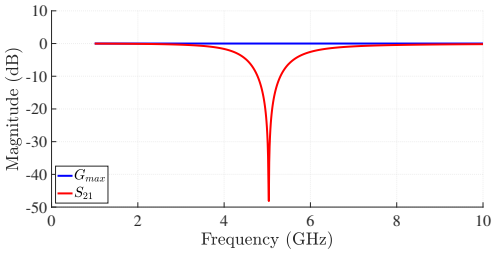
(c) Lossy Parallel LC



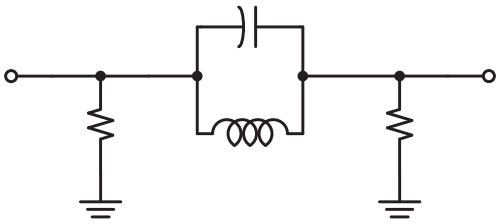
(d) Lossy Parallel LC performance



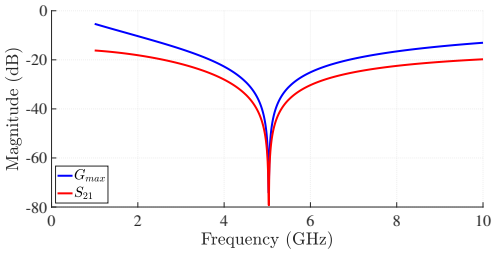
(e) Series LC



(f) Series LC performance



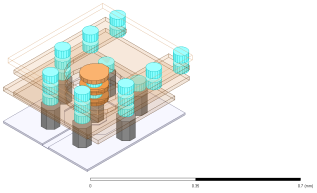
(g) Lossy Series LC



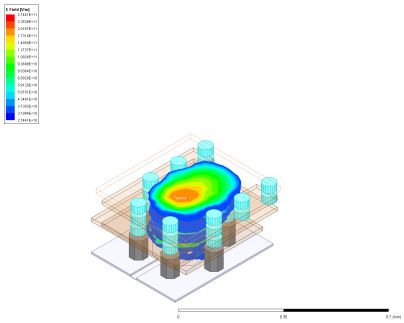
(h) Lossy Series LC performance

Figure 4.14: Difference between G_{max} and S_{21} in different lumped structures with $R = 10\Omega$, $L = 1\text{nH}$, $C = 1\text{pF}$

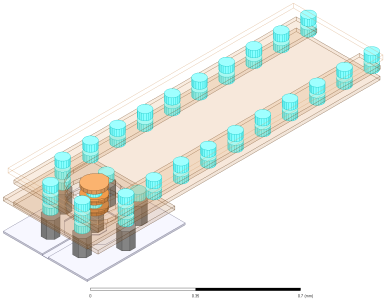
cross the cut-off frequency of the waveguide. For this reason, this transition structure should not be used beyond the TE cut-off frequency of the waveguide. To prove this theory, several different stripline lengths are simulated (Fig. 4.16). The simulation results (Fig. 4.17b) prove that the frequency of the notch varies with the length of the line. Moreover, multiple resonant modes can cause multiple notches. However, all of these notches persist above 300GHz and have almost no effect on the performance of the transition below 200GHz (Fig. 4.17a).



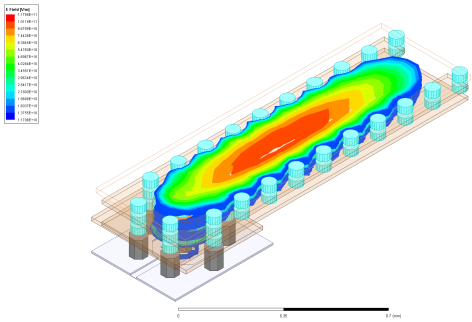
(a) Short Line



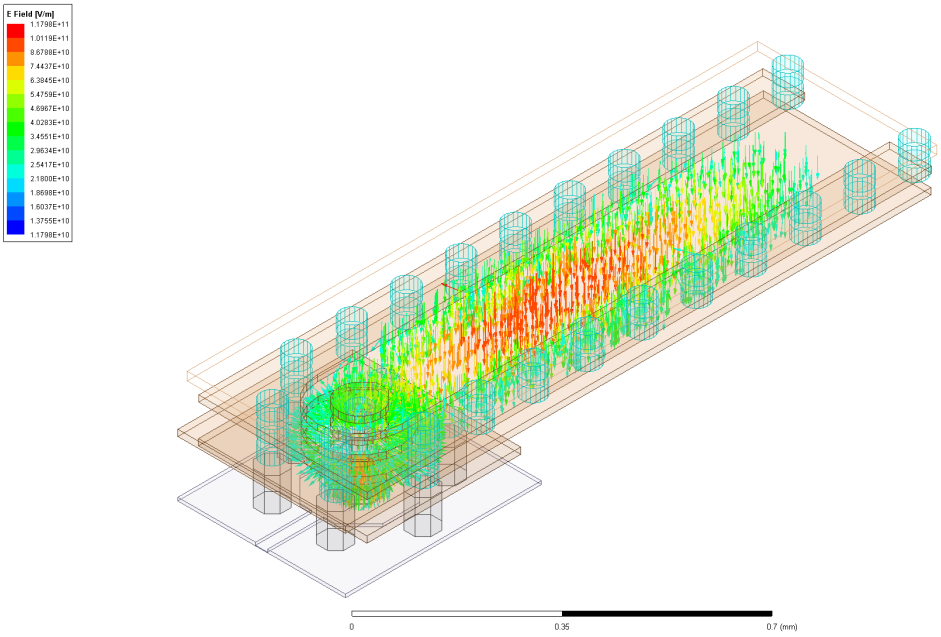
(b) E-field magnitude of the short line Eigen mode



(c) Long Line



(d) E-field magnitude of the long line Eigen mode



(e) E-field vector of the long line Eigen mode

Figure 4.15: Eigenmode simulation of resonant modes with different stripline length

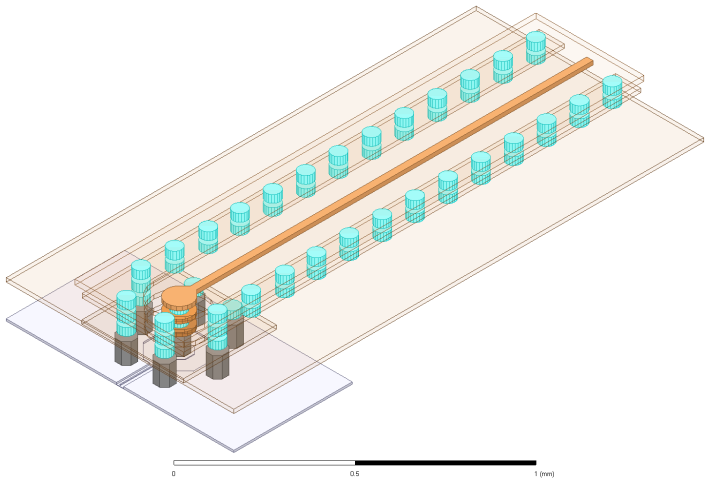
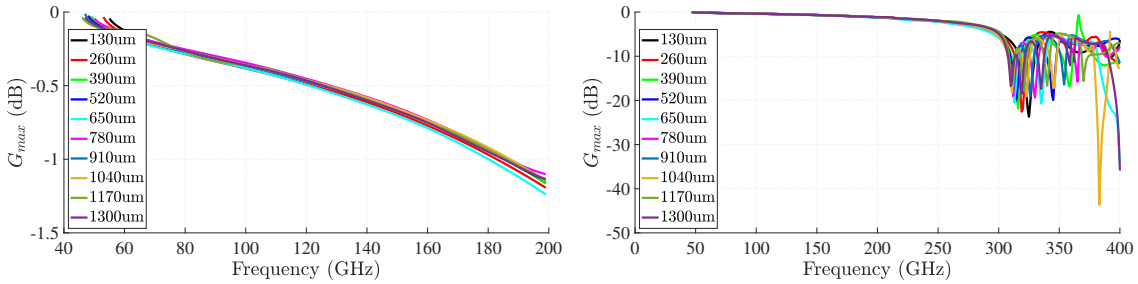


Figure 4.16: Long striplines are studied for the effects of cavity resonance



(a) Sub 200GHz (b) DC to 400GHz

Figure 4.17: G_{max} of the stripline transition when the length of the stripline extension is varied

4.4 Final Pad Structure

Given the advantages of the stripline transition over its counterparts, it was chosen for millimeter-wave I/Os. Below 300GHz, the transition can be modeled with two capacitors and a series transmission line representing the pad capacitance, the effective delay, and the characteristic impedance of the microvias from the stripline opening to the chip, as shown in Fig. 4.18. Note that the area inside the ground cage on the silicon is wasted if the matching network is implemented outside the pad area. Moreover, the access line can degrade the bandwidth and loss of the network (Fig. 4.19). Therefore, the matching network is implemented inside the ground cage. It consists of two symmetrical transmission lines (Fig. 4.20a), whose characteristic impedance and length are calculated to obtain a matched impedance (Fig. 4.20b).

The performance of the final design is simulated and shown in Fig. 4.21 and summarized

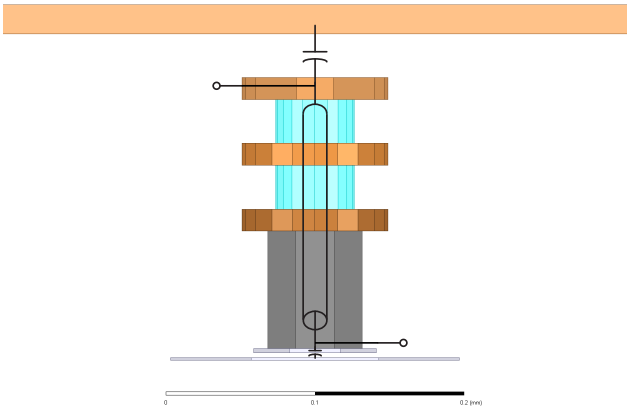


Figure 4.18: Lumped model of the transition below the stripline cut-off frequency

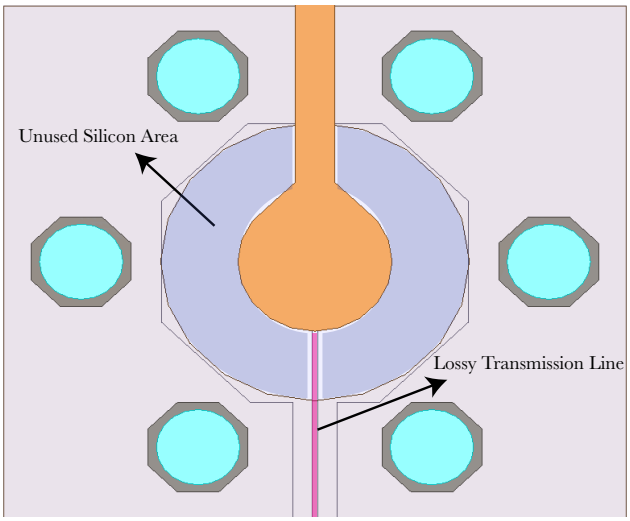
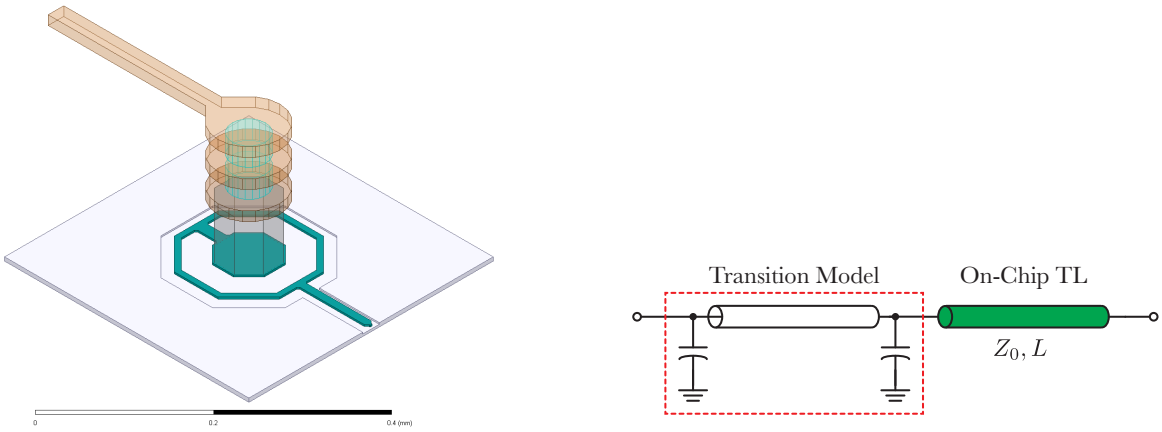


Figure 4.19: Wasted silicon area and additional losses due to the access line

in Table. 4.1. Table. 4.2 compares the performance obtained here with several other published papers.



(a) Physical Model (b) Lumped Model

Figure 4.20: The final design of the transition with a suitable matching network

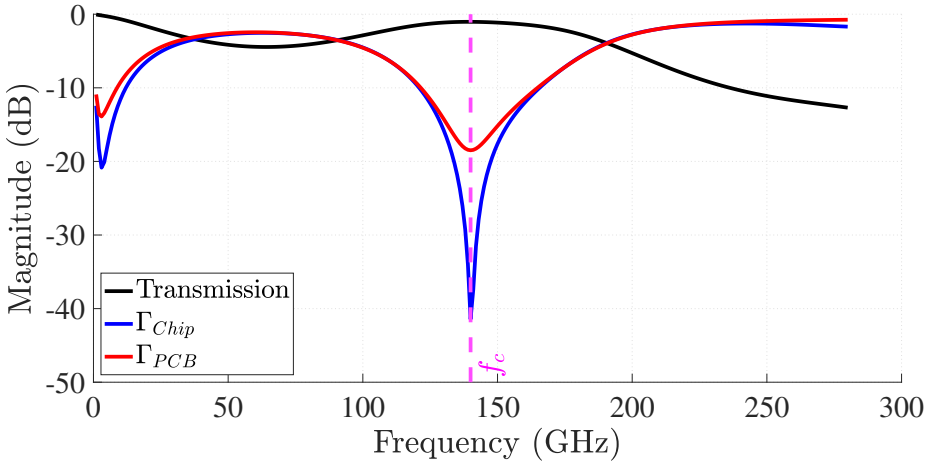


Figure 4.21: Performance of the final design

Parameter	Value
Center Frequency	140GHz
Insertion Loss	−1dB
3dB Bandwidth	85GHz
−10dB Reflection Bandwidth	43GHz

Table 4.1: Performance of the final design

Ref.	Package	Interconnect	Size	Pad Pitch	Frequency	Transition Loss
[49]	RO4350	Copper Pillar	-	-	130GHz	3dB
[50]	Astra MT77	-	-	-	145GHz	2.5dB
[51]	LTCC GL771	Copper Pillar	30μm	175μm	135GHz	1.1dB
[33]	Megtron 6	Solder Bump	-	250μm	115GHz	-
[52]	IPD carrier	Gold Bump	65μm	170μm	163GHz	2.8dB
This work	ABF GL102	Solder Bump	75μm	150μm	140GHz	1dB

Table 4.2: Summary of performance and comparison with the state-of-the-art

Chapter 5

Package-to-Package Transition

5.1 Introduction

Millimeter-wave and sub-THz systems offer unique applications for communication systems because higher total bandwidth and higher data rates can be achieved with a fixed fractional bandwidth [20]. However, due to excessive path loss in this frequency range, such systems must generate higher output power at the transmitter and achieve higher gain at the receiver. A phased array architecture (with N elements) can increase performance by relaxing the requirements on each element in the array and using the effective array gain. Given the enormous number of elements required to achieve high array gain [53, 54], the use of on-chip antennas is not feasible due to the cost of antenna area on semiconductors. Therefore, it makes sense to leave the antennas on the package and optimize the package materials and technology for higher radiation efficiency.

The decades of innovation and scaling of CMOS technology [55] makes it the first choice for designers when it comes to array processing. Although digital circuits have benefited dramatically from technological improvements, the analog and RF performance of CMOS has remained relatively similar over the past decade at $f_{max} \approx 300-400$GHz [56]. This makes CMOS extremely inefficient for sub-THz applications and encourages the coexistence of (III /V) compound semiconductors such as GaN or InP to boost performance. A package capable of carrying millimeter-wave and sub-THz signals with minimal insertion and radiation losses is needed. Also, high resolution and fine pitch are required to connect as many signals as possible to the chip with minimal reflection losses.

Thermal considerations are another aspect of sub-THz package design. The elements of a phased array are typically spaced $\lambda_0/2$ apart to minimize side lobes and mutual antenna

coupling. This means that as frequency increases, so does the heat flux [1], and the package should have excellent heat dissipation for reliable performance.

A single package that meets all the above requirements is expensive. This means that a modular package (as shown in Fig. 5.1) can optimize the cost and performance of millimeter-wave and sub-THz systems. However, a major practical problem is the transition of high-frequency signals between packages. Current low-cost solutions such as wire bonds and C4 bumps have low reliability for massive array implementation, or their resolution is insufficient to realize a low-loss transition due to reflections. This chapter proposes a new inter-package interconnect architecture based on guided inter-package radiation using mature and low-cost Ball Grid Arrays (BGA). Compared to other low-cost solutions, the proposed solution achieves higher bandwidth with lower insertion loss, while the lithography and alignment requirements are much more relaxed.

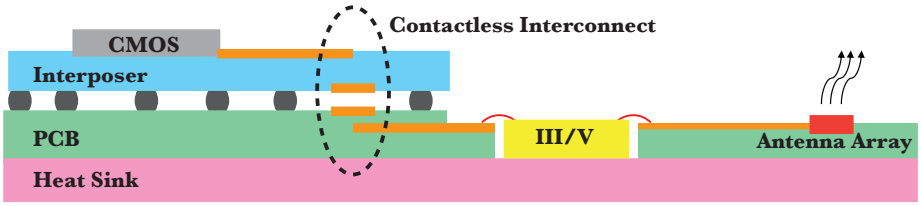


Figure 5.1: Proposed millimeter-wave phased array packaging solution with integrated III/V semiconductor

5.2 Design Principles

Proximity interconnects based on capacitive or inductive coupling have been explored for various applications, such as when isolation (thermal or electrical) is required [57]or when transceivers cannot be physically connected. In such systems, the receiver is located in the reactive near-field region of the transmitter. While this method works very well at lower frequencies, it is not readily possible to place transceivers in each other's reactive near-field in the millimeter-wave and sub-THz frequency range. For example, to transmit a signal with a frequency of 150GHz between two packages, their distance should be less than $200\mu m$, which requires good alignment during fabrication.

On the other hand, transmitting signals between two antennas in the far-field (Fraunhofer zone) is more common. This is how most conventional radio receivers operate. Far-field transmission, while simple, involves significant path loss, making it impractical for interconnects.

[1] Neglecting the lower device efficiency at higher frequencies.

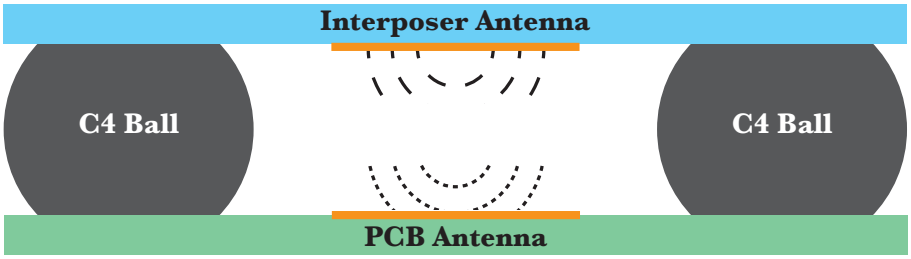


Figure 5.2: Millimeter-wave contactless inter-package interconnect based on guided radiation

When the distance between antennas is comparable to the wavelength ($d \approx 0.2\lambda...2\lambda$), the transceivers are in each other's Fresnel zone (radiative near-field). It is better to balance the two zones, as lower insertion loss can be achieved without stringent manufacturing requirements. However, electromagnetic fields tend to change rapidly with distance in this region. This effect can be modeled as the superposition of multiple propagation modes with different phase velocities. Perfect transmission occurs when all transmitter modes (having propagated through the channel at their velocities) match the corresponding receiver modes, or when the reflections of the different modes cancel each other out. This approach, while theoretically possible, requires strict alignment and precision in fabrication and usually has a narrow bandwidth.

The channel can be designed to have a preferred propagation mode to reduce sensitivity to distance. In this case, the channel rejects unwanted modes and allows only a single mode. Since the channel enforces modal purity, variations in the distance between antennas during fabrication only change the phase delay through the channel.

It can be compared to the performance of single-mode waveguides [58]. Waveguides are usually designed to have modal purity for an infinitely long channel. However, the transition distance between packages shown in Fig. 5.1 is generally about $d \approx 1$mm or less. Therefore, instead of a standard waveguide, a pseudo-waveguide can be designed to operate with only one dominant mode over a given channel length, which is far more relaxed than a conventional waveguide design.

Fig. 5.2 shows the principles of the proposed idea. For an inter-package interconnect, antennas on each package face each other, surrounded by BGA balls. These balls will shield the radiation to minimize leakage and insertion loss while rejecting undesired modes.

5.3 Design Considerations

Fig. 5.3 shows an example cross-section of two packages mounted on top of each other with a Ball Grid Array (BGA). There are several things to note here:

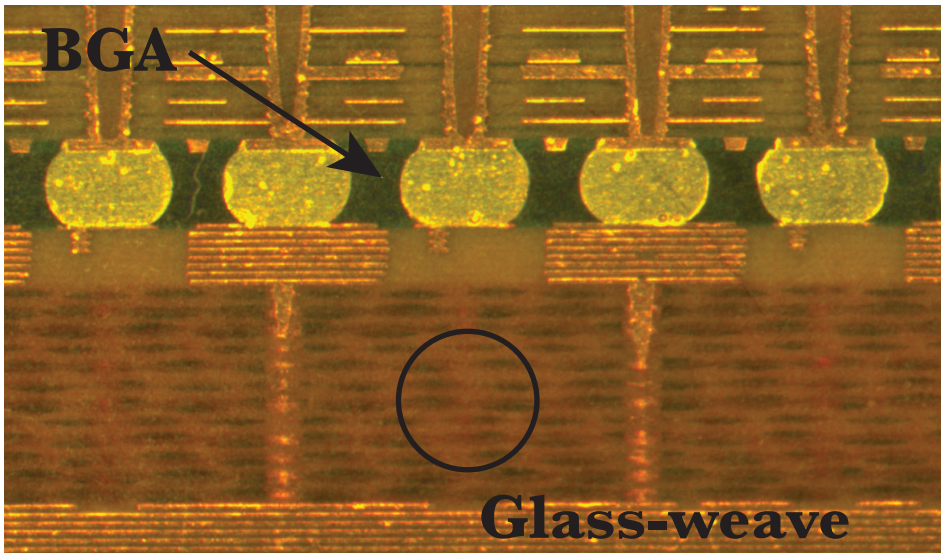


Figure 5.3: Cross section of two packages attached to each other using BGA balls

- With inexpensive BGAs, relatively good alignment can be achieved. Also, almost the same pattern repeats at the interface, indicating that reliable, predictable, and reproducible performance can be expected.
- The exact shape and curvature of the solder balls may vary. Therefore, the design should be such that the performance is less sensitive to the precise diameter of the balls, for example, by increasing the distance between the balls. In this case, the solder balls can be modeled as cylinders.
- Glass-weave may adversely impact the performance when the structure is much smaller than the periodicity of the woven structure. However, if the dimensions are chosen large enough, the radiators will see an average dielectric constant.

As mentioned earlier, the larger the structure is, the less sensitive it is to manufacturing variations and process nonidealities. However, the further away the solder balls are, the less shielding can be expected. Therefore, how much shielding can be expected from the BGA is unknown. To answer this question, two different scenarios for an incident wave are considered (Fig. 5.5):

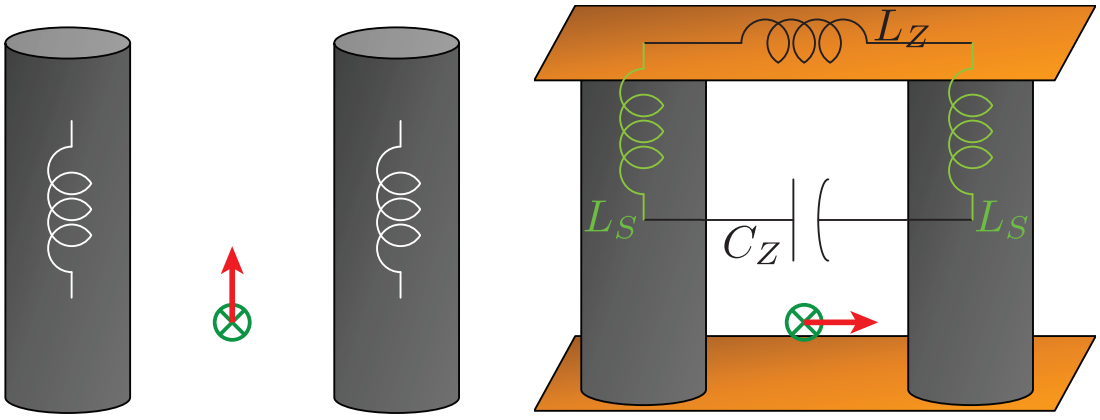
1. E-field parallel to the solder balls (Fig. 5.4a): In this case, the shielding is achieved by the induced current in the solder balls (which are modeled as cylinders), leading to an inductive reflection of the incident wave. Intuitively, as the ball pitch increases (for a fixed ball diameter), higher leakage should be expected since the incident wave is less coupled to the BGA balls. Moreover, the shielding performance is independent of the height of the balls (which determines the length of the pseudo-waveguide channel).

2. E-field perpendicular to the solder balls (Fig. 5.4b): In this case, individual BGA balls cannot provide sufficient shielding because the induced current is immediately interrupted by the discontinuity of the ball grid array. In this case, there is a redistribution of electric charge on the ball, indicating a (tiny) capacitive reflection of the incident wave. However, suppose that the solder balls are short-circuited by the interposer or the metal planes of the PCB. In this case, the induced charges cause an electric current to flow through the planes, and consequently, inductive reflection is expected. Increasing the ball pitch decreases the shielding performance for a fixed ball diameter since a smaller electric charge is initially induced on the balls. As the height of the ball increases (assuming its diameter can be kept constant), the shielding decreases as the same induced charges on the balls experience a higher series inductance before reaching the metal planes.

   Since the reflection depends on the LC loop formed by the BGA and metal planes, there is a resonant frequency at which no shielding is expected. Assuming a simple LC model

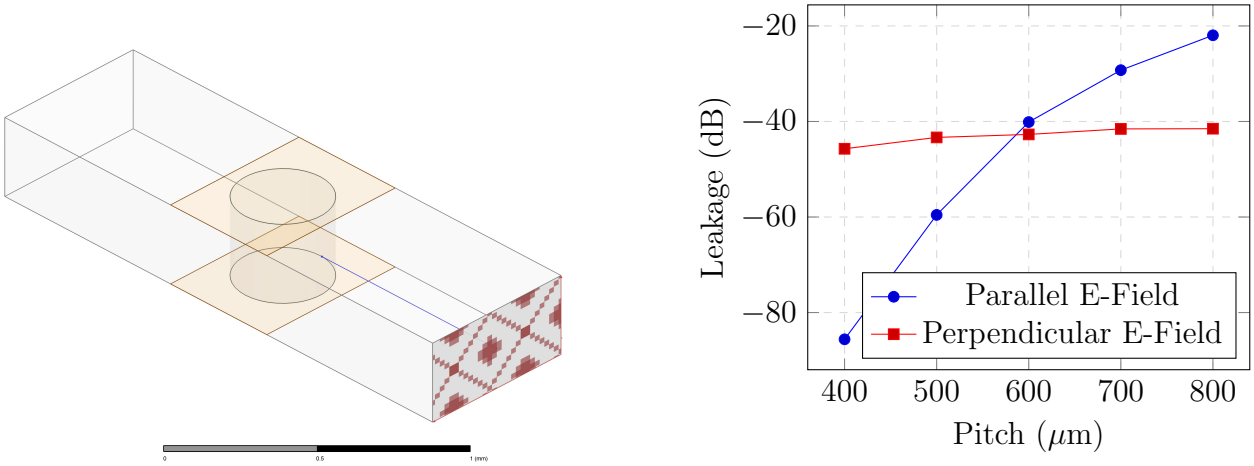
$$\omega_0{}^{-2} = (C_Z)(2L_S + L_Z) \tag{5.1}$$

The effectiveness of the BGAs for shielding was verified using the full-wave simulation software ANSYS HFSS. In this simulation (Fig. 5.5a), unit cells with slave/master boundary conditions are used. A ball diameter of 350μm is chosen, and the ball height is assumed to be equal to the ball diameter. Fig. 5.5b shows the simulation results at 140GHz. They show that the BGA can effectively reflect incident waves and mimic a solid metallic plane for the frequency range of interest.



(a) Parallel E-field (b) Perpendicular E-field

Figure 5.4: The lumped circuit model seen by an incoming wave with specific E-polarization



(a) Unit cells simulated by ANSYS HFSS

(b) Simulation results

Figure 5.5: Simulation results for the leakage of an incident wave with the frequency of 140GHz with different E-field polarization upon arrival to a shorted BGA with ball diameter of $350\mu m$

5.4 Prototype Design and Measurement Results

A contactless interconnect was developed as a proof of concept. In this section, various aspects of the design methodology are explained.

Interposer Technology

The interposer used here is made of organic materials. It consists of six build-up layers (ABF GL102) with a total thickness of $300\mu m$ symmetrically attached to a $400\mu m$- thick core layer (MCL-E-705G) for mechanical support, as shown in Fig. 5.6. The fabrication capabilities allow the microvias of the build-up layers to have a spacing of only $100\mu m$, while the plated through holes in the core layer have a minimum spacing of $300\mu m$.

Channel Design Trade-offs

Assuming that solder balls can provide sufficient shielding, cylindrical balls are connected from the outer sides to form a pseudo-waveguide, as shown in Fig. 5.7. Then a modal simulation is performed to find the propagation modes. Different dimensions of the pseudo-waveguide (by changing the ball pitch, the ball diameter, and the number of balls in each row) are investigated. As a compromise between modal purity, the characteristic impedance of the desired mode, frequency dispersion, and fabrication capabilities, the structure of Fig. 5.7

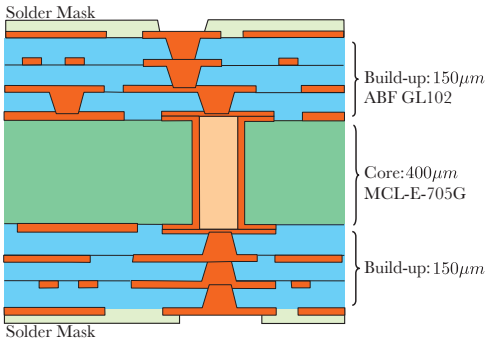


Figure 5.6: Organic interposer technology

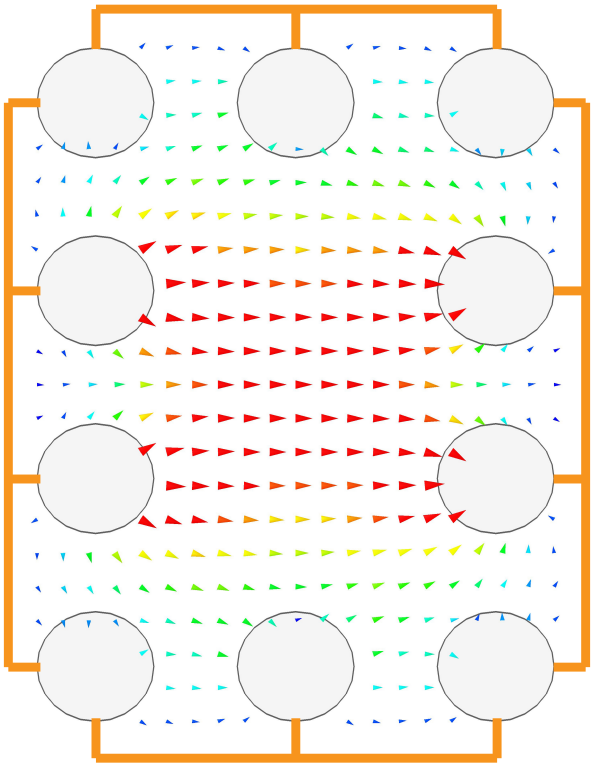


Figure 5.7: Pseudo-waveguide port definition

with a ball diameter of 350μm and a ball pitch of 600μm is chosen. Once the dimensions of the structure are determined, the shielding performance is simulated and verified using the technique described in the previous section.

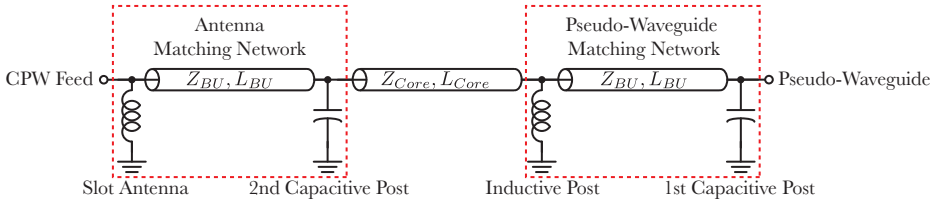


Figure 5.8: Lumped model of the distributed matching network

Antenna Design with Distributed Matching Network

As shown in Fig. 5.1, the millimeter-wave contactless interconnect is fed from one side of the interposer (slot- antenna fed from a CPW line), while the receiver (on the PCB) is located on the other side of the interposer. Since the electrical length of the core and build-up layers is comparable to the wavelength, a lumped matching network at the excitation site leads to low bandwidth and high insertion loss. Therefore, a distributed matching network is used.

To facilitate the design of the matching network and avoid exhaustive electromagnetic simulations, the matching network is first divided into two parts as explained in Fig. 5.8. The first part matches the impedance of the pseudo-waveguide to the impedance of the wave in the core layer [2]. The second part is about matching the slot antenna to the impedance of the waves in the core layer. The reason for this decision is that the core layer is thick (with an electrical length of $\approx 150°$), and has a higher dielectric loss than the build-up layers. Therefore, any reflection within the core results in higher insertion loss and lower bandwidth. Once the matching network is roughly calculated, the correct values (for the size of the inductive and capacitive posts) are entered into the simulator. After running the optimization engine to fine-tune the entire structure, we found that the initial calculated values were close to optimal. The final design and exploded view are shown in Fig. 5.9 and Fig. 5.10, respectively.

Prototype Performance

A prototype is simulated and fabricated (Fig. 5.11) to verify the proposed solution and design methodology. It consists of a back-to-back connection of two millimeter-wave contactless interconnects which (Fig. 5.12). The simulation and measurement results are shown in Fig. 5.13. It is observed that 20GHz of -10dB reflection bandwidth is achievable with 4dB insertion loss for a back-to-back structure (2dB loss for each leg). The additional insertion loss in the measurement compared to the simulation results is likely due to the surface roughness of the copper.

[2]The core dielectric and the plated through holes together form another pseudo- waveguide.

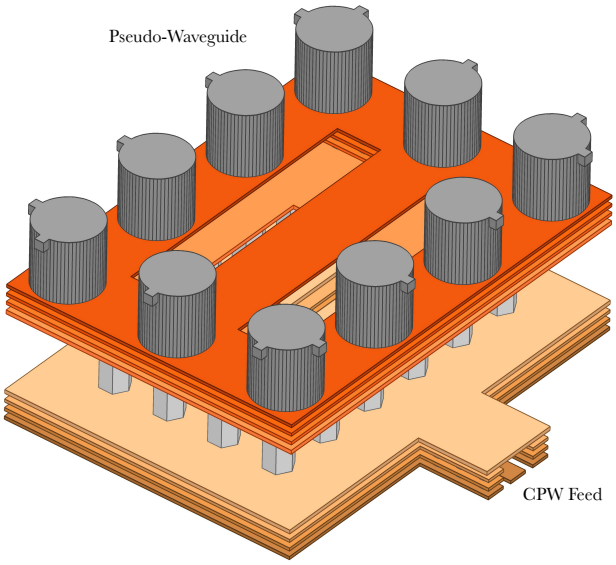


Figure 5.9: Interposer antenna

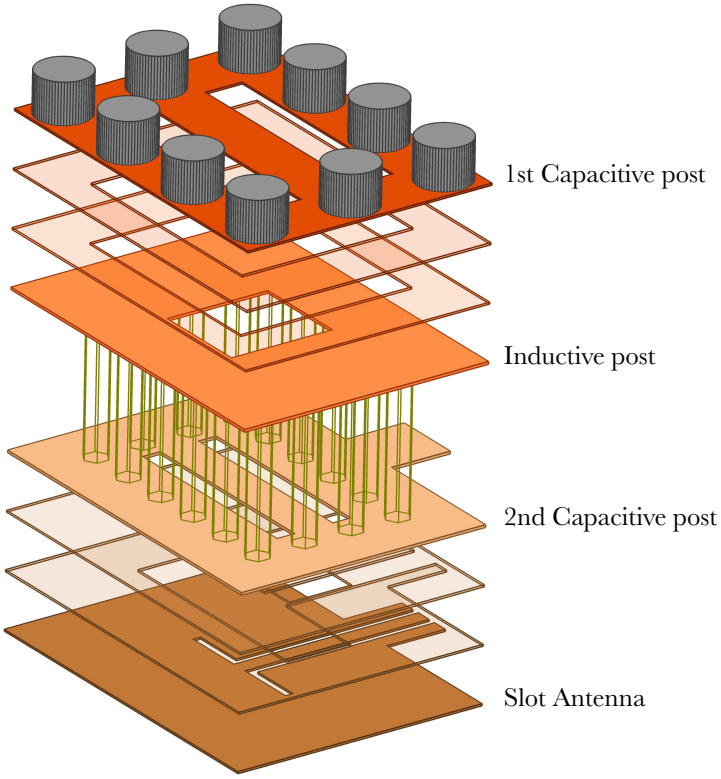
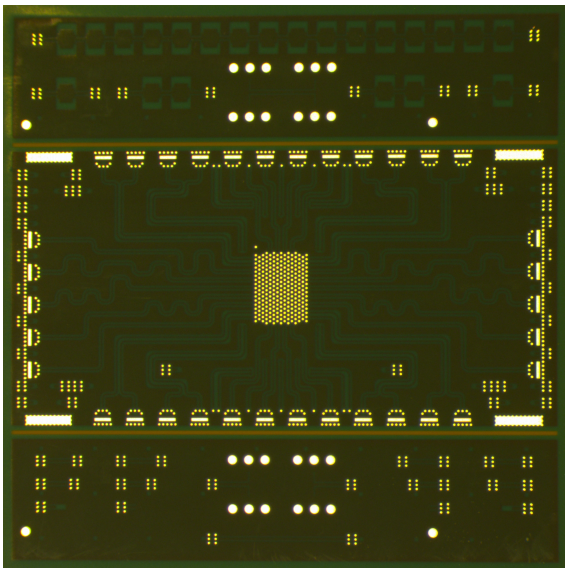
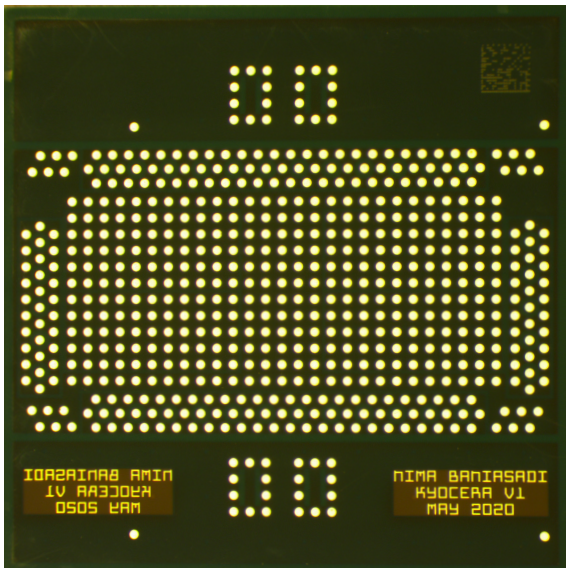


Figure 5.10: Exploded view of the antenna (microvias not shown)

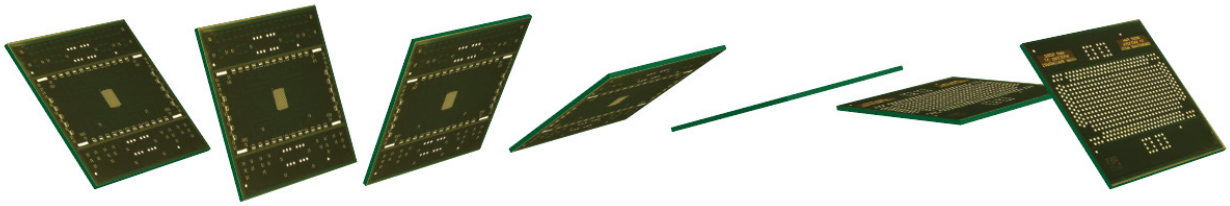


(a) Top view

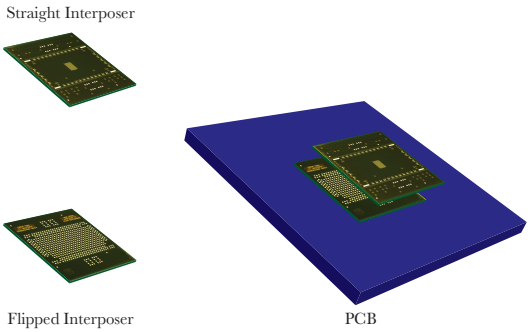


(b) Bottom view

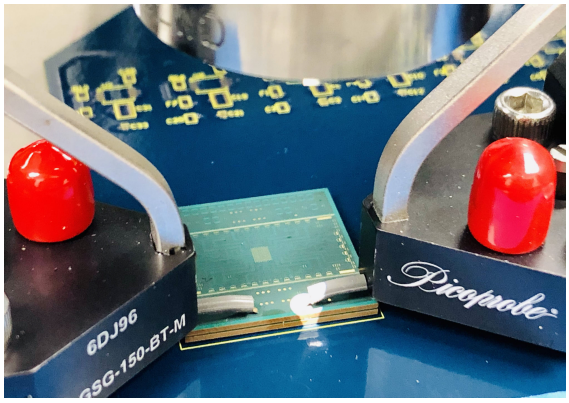
Figure 5.11: Fabricated millimeter-wave contactless interconnect



(a) Flipping for attaching



(b) Back-to-back attachment



(c) Measurement

Figure 5.12: Back-to-back millimeter-wave contactless interconnect

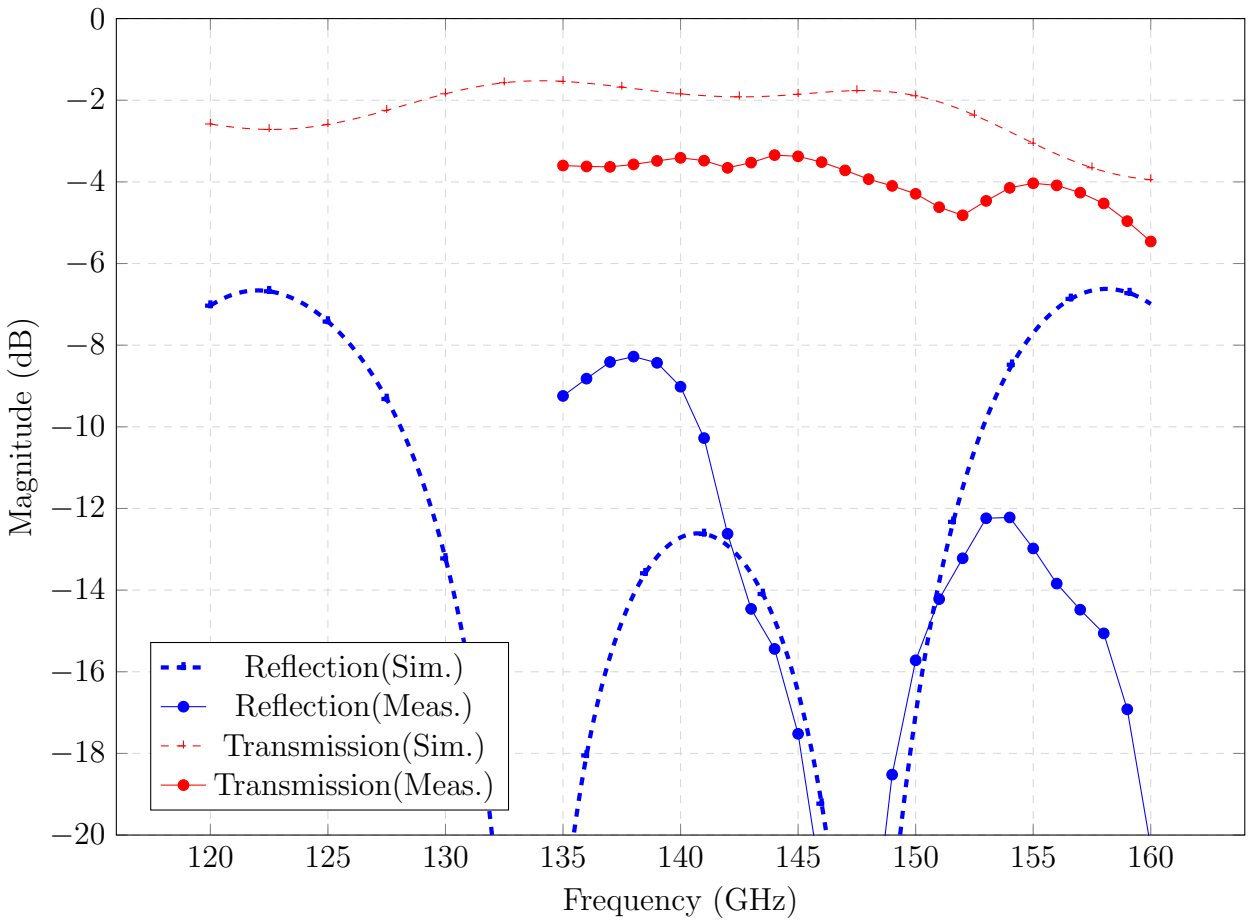


Figure 5.13: Simulation and measurement results for a back-to-back millimeter-wave contactless interconnect

5.5 Conclusion

The millimeter-wave contactless interconnect based on guided radiation is proposed as a new method for inter-package routing. Design methods and guidelines are explained to obtain an estimated performance before performing detailed electromagnetic simulations. A distributed matching network is also proposed to achieve high bandwidth and low insertion loss. Full-wave electromagnetic simulations verify all proposed ideas and methods. Finally, the prototype is fabricated and measured. The measurements agree well with the simulation results.

Chapter 6

Conclusion

6.1 Thesis Summary

The next generation of mobile communications requires cost-effective solutions to increase the capacity of the cellular network. Millimeter-wave carrier frequencies enable high-speed links in a lightly licensed portion of the spectrum. However, CMOS process scaling is no longer as effective as it was in the past to enable high-speed applications. CMOS scaling can degrade device output power at high frequencies.

Noise measure is considered as a performance metric that combines the power gain with the minimum noise figure given the limited power gain of devices operating close to their activity limit. Enlightening examples allow the reader to grasp the mathematical framework intuitively. The use case of active baluns is explored using noise measure theory, and optimal working conditions are investigated.

The design of a wideband receiver at 140GHz is discussed. Several different techniques are proposed to improve the performance of the receiver chain compared to the state-of-the-art. All of these techniques are mathematically proven, and tradeoffs are explored. These techniques, such as transformer equivalents, active baluns, and optimal matching networks, can be readily implemented in commercial ICs to improve performance.

Finally, cost-effective packaging solutions for millimeter-wave applications have been explored. Note that much of the published work was either measured with probes or packaged with on-chip antennas, possibly with integrated silicon lenses. As with commercial applications, the transition from chip to package and between packages has been investigated. It has been shown that currently available low-cost package options can meet millimeter-wave requirements.

6.2 Future Directions

As shown in this work, conjugate matching does not provide optimal performance. Therefore, transmission is optimized, and matching networks are designed to achieve optimal transmission. Although low-k transformers have been used extensively in this work, they were not intentionally designed for high bandwidth. In other words, the high bandwidth is just a byproduct of using low-k transformers. The simulation results show that a combination of LC ladder networks with transformers can deliver the maximum transmission while intentionally maximizing the system's bandwidth.

Another avenue of research is to investigate the performance of common-gate amplifiers. Note that there is no difference between the two amplifiers from the noise measure point of view. However, the power gain of common-gate amplifiers is lower than that of common-source amplifiers. On the other hand, the insertion loss over the matching network is lower as expected due to the lower input quality factor. Thus, if the insertion loss of the matching network is significant, a common-gate stage may be superior to a common-source counterpart. Also, since the input signal is not connected to the gate, it is easy to use a double-sided contact with minimal parasitic capacitance.

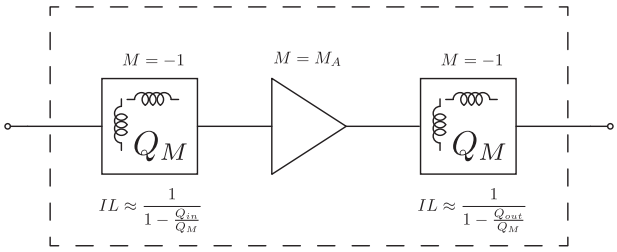


Figure 6.1: Estimating the noise measure of an amplifier including the inserion loss of the matching networks

Finally, given the equations for the minimum noise measure and the insertion loss of the matching network, you can combine the two to derive the minimum noise measure of an amplifier with its matching network (Fig. 6.1).

Bibliography

[1] K. Pahlavan and P. Krishnamurthy, "Evolution and impact of wi-fi technology and applications: A historical perspective," *International Journal of Wireless Information Networks*, vol. 28, no. 1, pp. 3–19, Mar 2021. [Online]. Available: https://doi.org/10.1007/s10776-020-00501-8

[2] "Ieee standard for safety levels with respect to human exposure to electric, magnetic, and electromagnetic fields, 0 hz to 300 ghz," *IEEE Std C95.1-2019 (Revision of IEEE Std C95.1-2005/ Incorporates IEEE Std C95.1-2019/Cor 1-2019)*, pp. 1–312, 2019.

[3] [Online]. Available: https://gems.ece.gatech.edu/PA_survey.html

[4] T. Naveh, "Mobile backhaul: Fiber vs. microwave." [Online]. Available: https://www.winncom.com/images/stories/Ceragon_Mobile_Backhaul_Fiber_Microwave_WP.pdf

[5] M. Gamage, "Wireless backhaul for 5g." [Online]. Available: https://www.pta.gov.pk/media/satrc/02-170817_9.pdf

[6] J. Saunders and N. Marshall, "Mobile backhaul options - gsma." [Online]. Available: https://www.gsma.com/spectrum/wp-content/uploads/2019/04/Mobile-Backhaul-Options.pdf

[7] "Ericsson microwave outlook 2020." [Online]. Available: https://wcm.ericsson.net/4a811d/assets/local/reports-papers/microwave-outlook/2020/2020-ericsson-microwave-outlook-report-digital.pdf

[8] D. M. Pozar, *Microwave engineering*, 3rd ed. Hoboken, NJ: J. Wiley, 2005.

[9] [Online]. Available: https://docs.fcc.gov/public/attachments/FCC-19-126A1.pdf

[10] [Online]. Available: https://docs.fcc.gov/public/attachments/DOC-348982A1.pdf

[11] C. A. Balanis, *Antenna theory : analysis and design*, fourth edition. ed. Hoboken, New Jersey: Wiley, 2016 - 2016.

[12] G. Masetti, M. Severi, and S. Solmi, "Modeling of carrier mobility against carrier concentration in arsenic-, phosphorus-, and boron-doped silicon," *IEEE Transactions on Electron Devices*, vol. 30, no. 7, pp. 764–769, 1983.

[13] E. Johnson, "Physical limitations on frequency and power parameters of transistors," in *1958 IRE International Convention Record*, vol. 13, 1965, pp. 27–34.

[14] J. McPherson, J. Kim, A. Shanware, H. Mogul, and J. Rodriguez, "Proposed universal relationship between dielectric breakdown and dielectric constant," in *Digest. International Electron Devices Meeting,*, 2002, pp. 633–636.

[15] C. Enz, "An mos transistor model for rf ic design valid in all regions of operation," *IEEE Transactions on Microwave Theory and Techniques*, vol. 50, no. 1, pp. 342–359, 2002.

[16] C.-H. Choi, J.-S. Goo, Z. Yu, and R. Dutton, "Shallow source/drain extension effects on external resistance in sub-0.1 /spl mu/m mosfets," *IEEE Transactions on Electron Devices*, vol. 47, no. 3, pp. 655–658, 2000.

[17] S.-D. Kim, C.-M. Park, and J. Woo, "Advanced model and analysis of series resistance for cmos scaling into nanometer regime. i. theoretical derivation," *IEEE Transactions on Electron Devices*, vol. 49, no. 3, pp. 457–466, 2002.

[18] S. Thompson *et al.*, "A 90-nm logic technology featuring strained-silicon," *IEEE Transactions on Electron Devices*, vol. 51, no. 11, pp. 1790–1797, 2004.

[19] S. Li, Z. Zhang, B. Rupakula, and G. M. Rebeiz, "An eight-element 140-ghz wafer-scale if beamforming phased-array receiver with 64-qam operation in cmos rfsoi," *IEEE Journal of Solid-State Circuits*, pp. 1–1, 2021.

[20] A. Townley *et al.*, "A Fully Integrated, Dual Channel, Flip Chip Packaged 113 GHz Transceiver in 28nm CMOS supporting an 80 Gb/s Wireless Link," *Proceedings of the Custom Integrated Circuits Conference*, vol. 2020-March, pp. 113–116, 2020.

[21] H. A. Haus and R. B. Adler, "Optimum noise performance of linear amplifiers," *Proceedings of the IRE*, vol. 46, no. 8, pp. 1517–1533, 1958.

[22] ——, *Impedance Formulation of the Characteristic-Noise Matrix*, 1959, pp. 19–27.

[23] M. Pospieszalski, "Modeling of noise parameters of mesfets and modfets and their frequency and temperature dependence," *IEEE Transactions on Microwave Theory and Techniques*, vol. 37, no. 9, pp. 1340–1350, 1989.

[24] S. Mason, "Power gain in feedback amplifier," *Transactions of the IRE Professional Group on Circuit Theory*, vol. CT-1, no. 2, pp. 20–25, 1954.

[25] A. M. Niknejad, *Electromagnetics for High-Speed Analog and Digital Communication Circuits*. Cambridge University Press, 2007.

[26] A. Singhakowinta and A. Boothroyd, "On linear two-port amplifiers," *IEEE Transactions on Circuit Theory*, vol. 11, no. 1, pp. 169–169, 1964.

[27] T. Grasser, Ed., *Noise in Nanoscale Semiconductor Devices*. Springer International Publishing, 2020, ch. 6. [Online]. Available: https://doi.org/10.1007%2F978-3-030-37500-3

[28] C. Poole and I. Darwazeh, *Microwave Active Circuit Analysis and Design*. San Diego, CA, USA: Elsevier Science, 2015.

[29] S. Mohan, M. del Mar Hershenson, S. Boyd, and T. Lee, "Simple accurate expressions for planar spiral inductances," *IEEE Journal of Solid-State Circuits*, vol. 34, no. 10, pp. 1419–1424, 1999.

[30] D. Dubuc *et al.*, "High quality factor and high self-resonant frequency monolithic inductor for millimeter-wave Si-based IC's," *IEEE MTT-S International Microwave Symposium Digest*, vol. 1, pp. 193–196, 2002.

[31] P.-O. Leine, "On the power gain of unilaterized active networks," *IRE Transactions on Circuit Theory*, vol. 8, no. 3, pp. 357–358, Sep. 1961.

[32] A. Mazzanti and A. Bevilacqua, "Second-order equivalent circuits for the design of doubly-tuned transformer matching networks," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 12, pp. 4157–4168, Dec 2018.

[33] A. Townley, "Broadband mm-wave transceivers for sensing and communication," Ph.D. dissertation, EECS Department, University of California, Berkeley, May 2020. [Online]. Available: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2020/EECS-2020-25.html

[34] S. V. Thyagarajan, "Millimeter-wave/terahertz circuits and systems for wireless communication," Ph.D. dissertation, EECS Department, University of California, Berkeley, May 2016. [Online]. Available: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2016/EECS-2016-22.html

[35] S. Krishnamurthy and A. Niknejad, "Fanout optimization for an inductorless broadband variable gain cherry-hooper amplifier," Master's thesis, EECS Department, University of California, Berkeley, May 2021. [Online]. Available: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-23.html

[36] *Broadband Circuits for Optical Fiber Communication*. John Wiley & Sons, Ltd, 2005, ch. 6, pp. 159–232. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471726400.ch6

[37] S. Ma, J. Lin, C. Ma, and H. Yu, "A 140 ghz transceiver for ¡tex¿4×4¡/tex¿ beamforming short-range communication in 65nm cmos," in *2019 IEEE Asia-Pacific Microwave Conference (APMC)*, 2019, pp. 1613–1615.

[38] S. Li and G. M. Rebeiz, "A 134-149 ghz if beamforming phased-array receiver channel with 6.4-7.5 db nf using cmos 45nm rfsoi," in *2020 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*, 2020, pp. 103–106.

[39] A. Simsek, S.-K. Kim, and M. J. Rodwell, "A 140 ghz mimo transceiver in 45 nm soi cmos," in *2018 IEEE BiCMOS and Compound Semiconductor Integrated Circuits and Technology Symposium (BCICTS)*, 2018, pp. 231–234.

[40] A. A. Farid, A. Simsek, A. S. H. Ahmed, and M. J. W. Rodwell, "A broadband direct conversion transmitter/receiver at d-band using cmos 22nm fdsoi," in *2019 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*, 2019, pp. 135–138.

[41] N. G. Weimann *et al.*, "Manufacturable low-cost flip-chip mounting technology for 300-500-GHz assemblies," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 7, no. 4, pp. 494–501, 2017.

[42] V. Valenta *et al.*, "Design and experimental evaluation of compensated bondwire interconnects above 100 GHz," *International Journal of Microwave and Wireless Technologies*, vol. 7, no. 3-4, pp. 261–270, 2015.

[43] W. Heinrich, "The flip-chip approach for millimeter-wave packaging," *IEEE Microwave Magazine*, vol. 6, no. 3, pp. 36–45, 2005.

[44] S. Monayakul *et al.*, "Flip-Chip Interconnects for 250 GHz Modules," *IEEE Microwave and Wireless Components Letters*, vol. 25, no. 6, pp. 358–360, 2015.

[45] A. Jentzsch and W. Heinrich, "Theory and measurements of flip-chip interconnects for frequencies up to 100 GHz," *IEEE Transactions on Microwave Theory and Techniques*, vol. 49, no. 5, pp. 871–878, 2001.

[46] D. Deslandes and K. Wu, "Integrated microstrip and rectangular waveguide in planar form," *IEEE Microwave and Wireless Components Letters*, vol. 11, no. 2, pp. 68–70, 2001.

[47] K. Wu, D. Deslandes, and Y. Cassivi, "The substrate integrated circuits - a new concept for high-frequency electronics and optoelectronics," in *6th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Service, 2003. TELSIKS 2003.*, vol. 1, 2003, pp. P–III.

[48] Y. Cassivi *et al.*, "Dispersion characteristics of substrate integrated rectangular waveguide," *IEEE Microwave and Wireless Components Letters*, vol. 12, no. 9, pp. 333–335, 2002.

[49] M. Sawaby, N. Dolatsha, B. Grave, C. Chen, and A. Arbabian, "A fully packaged 130-ghz qpsk transmitter with an integrated prbs generator," *IEEE Solid-State Circuits Letters*, vol. 1, no. 7, pp. 166–169, 2018.

[50] A. Simsek, A. S. H. Ahmed, A. A. Farid, U. Soylu, and M. J. W. Rodwell, "A 140ghz two-channel cmos transmitter using low-cost packaging technologies," in *2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, 2020, pp. 1–3.

[51] A. A. Farid, A. S. H. Ahmed, A. Simsek, and M. J. W. Rodwell, "A packaged 135ghz 22nm fd-soi transmitter on an ltcc carrier," in *2021 IEEE MTT-S International Microwave Symposium (IMS)*, 2021, pp. 713–716.

[52] C.-H. Li, W.-T. Hsieh, and T.-Y. Chiu, "A flip-chip-assembled w-band receiver in 90-nm cmos and ipd technologies," *IEEE Transactions on Microwave Theory and Techniques*, vol. 67, no. 4, pp. 1628–1639, 2019.

[53] S. Shahramian, M. J. Holyoak, A. Singh, and Y. Baeyens, "A fully integrated 384-element, 16-tile, w -band phased array with self-alignment and self-test," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 9, pp. 2419–2434, Sep. 2019.

[54] U. Kodak, B. Rupakula, S. Zihir, and G. M. Rebeiz, "60-ghz 64- and 256-element dual-polarized dual-beam wafer-scale phased-array transceivers with reticle-to-reticle stitching," *IEEE Transactions on Microwave Theory and Techniques*, vol. 68, no. 7, pp. 2745–2767, July 2020.

[55] M. T. Bohr and I. A. Young, "CMOS Scaling Trends and beyond," *IEEE Micro*, vol. 37, no. 6, pp. 20–29, 2017.

[56] W. Steyaert and P. Reynaert, "Layout optimizations for THz integrated circuit design in bulk nanometer CMOS," *2017 IEEE Compound Semiconductor Integrated Circuit Symposium, CSICS 2017*, vol. 2017-January, pp. 1–4, 2017.

[57] M. De Wit, S. Ooms, B. Philippe, Y. Zhang, and P. Reynaert, "Polymer microwave fibers: A new approach that blends wireline, optical, and wireless communication," *IEEE Microwave Magazine*, vol. 21, no. 1, pp. 51–66, 2020.

[58] J. W. Holloway, G. C. Dogiamis, and R. Han, "Innovations in terahertz interconnects: High-speed data transport over fully electrical terahertz waveguide links," *IEEE Microwave Magazine*, vol. 21, no. 1, pp. 35–50, 2020.