Active Microwave Circuits

Zoya Popović and Laila Marzall

Abstract These are notes for ECEN 5104, a graduate course at the University of Colorado, Boulder. The notes are developed as the class progresses during the Spring 2024 semester. The authors appreciate any suggestions and corrections contributed by students.

1 Review: Impedance Matching

1.1 Lumped-Element Matching

The best review of lumped-element matching is given in the paper by Rhea in two parts, posted on the web page. Please read the paper. The simplest lumped-element matching circuits use a single inductor and capacitor, such as shown in Fig. 1 for the case of a parallel C and series L. In principle, any load can be matched with one of these circuits if the elements are chosen correctly. First consider the circuit on the left in the figure. The series inductance adds a positive reactance $jX_L = j\omega L$, so with an inductor alone, only impedances of the form $z = 1–jx$ can be matched ($x \geq 0$). This corresponds to the bottom half of the $r = 1$ circle of the passive Smith chart. A capacitor is needed to bring the load impedance to this semi-circle. This can only be done if the load is either inside the $r = 1$ circle or above both the $g = 1$ and $r = 1$ circle circles. The strategy is to bring enough capacitive susceptance to bring the load onto the bottom half of the $r = 1$ circle, and then add inductance to bring the load into the center (Fig. 2).

Next consider the circuit on the right of Fig. 1. With capacitance alone, only loads of the form $y = 1–jb$ can be matched ($b \geq 0$). This is the top half of the $g = 1$ circle. An inductor wcan bring the load to this semi-circle, provided the load is inside the $g = 1$ circle or above both the $g = 1$ and $r = 1$ circle circles. The strategy in this case is to add enough inductance to bring the load onto the top half of the $g = 1$ circle and then add capacitance to bring the load to the center of the Smith chart. Fig. 2 shows this graphically.

1.2 Transmission-Line Stub Matching

Transmission-line stub matching is discussed in every textbook. Please review it or come ask me if you need review material. The one thing that you might want to keep

Zoya Popović
The University of Colorado, Boulder, CO, U.S.A., e-mail: zoya@colorado.edu

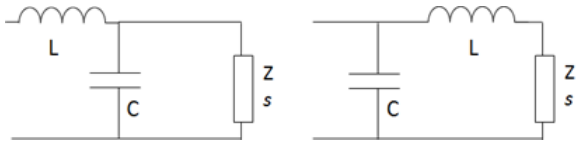


Fig. 1 *LC* matching networks that use a series inductance and parallel capacitance.

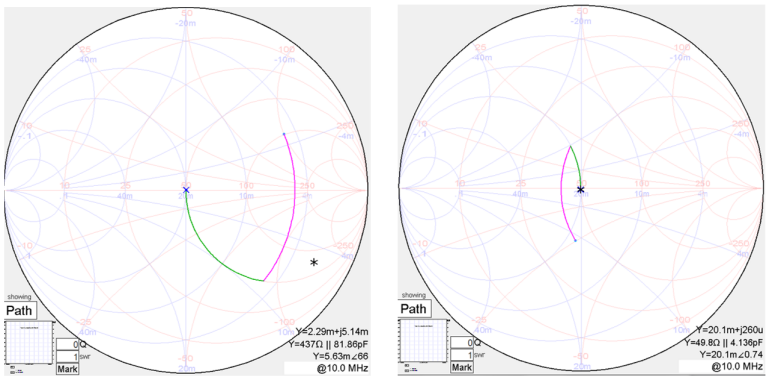


Fig. 2 *LC* matching on the Smith chart.

in mind is that the impedances of the lines and stubs do not need to be Z_0 like in textbooks. The reason for originally using only Z_0 lines is that people were doing this type of matching in coaxial lines and only one characteristic impedance was available. However, in microstrip, it is easy to get a range from 20 to 90 Ω, so this gives more options for the stub matching.

1.3 Quarter-Wave Line Match

A single quarter-wave long line is an impedance transformer with an input impedance equal to $Z_{in} = Z^2/Z_L$, where Z is the characteristic impedance of the line used for matching. For a lossless line, to make the input 50 Ω (or any real impedance), it is easy to see that only real loads $Z_L = R_L$ can be matched. To match a complex impedance using a quarter-wave line, one can add a length of 50-Ω (or other Z_0) line in order to make the load real and then add a quarter-wave match, Fig. 3.

1.4 Slug Matching

Cascaded sections of transmission lines with varying impedances that can also vary in position are referred to as slugs. This type of matching is used in coaxial circuits, and can be easily varied mechanically. One type of slug matching is shown in Fig. 4 where two dielectric cylinders can be moved up and down a section of air coaxial

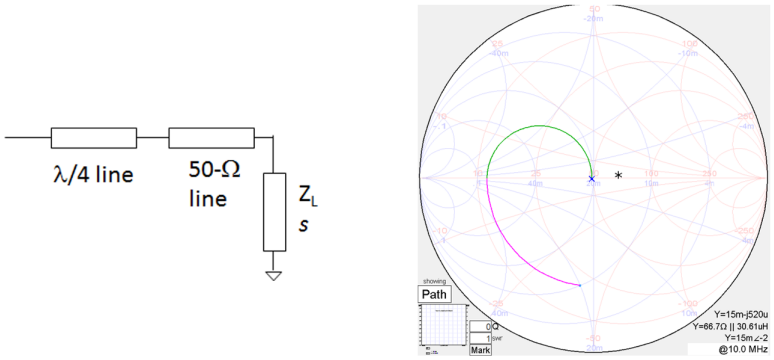


Fig. 3 Quarter-wave line matching on the Smith chart.

line. For a fixed load, they can then be glued into place after the load is matched. The impedance in the dielectric slug is reduced by from the impedance of the rest of the line, which is usually 50 Ω. The tuner covers the widest range when the electrical length of the slugs is 90° at the operating frequency.

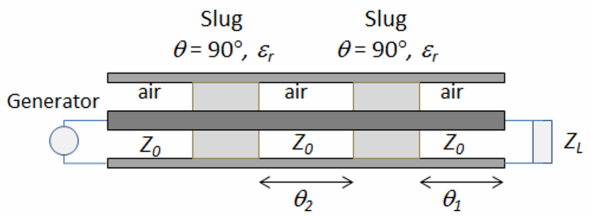


Fig. 4 Slug matching with two quarter-wave long dielectric slugs inserted in an air coaxial line.

First we can determine the region of the Smith chart that can be matched. Since we can make the length of the 50-Ω line in front of the load anything we want, the phase of the load does not matter. This means that the boundary of the region of the Smith chart that we can reach is a circle about the origin. The magnitude of the reflection coefficient that we match is determined by the spacing between the slugs. It varies from zero when the slugs touch each other up to its largest value when the separation is 90°. The maximum can be calculated by cascading quarter-wave section transformations. The impedance, looking back through the two slugs, is $50/\epsilon_r^2$. This means that we can match a reflection coefficient that satisfies

$$|s| \leq \frac{\epsilon_r^2 - 1}{\epsilon_r^2 + 1} \tag{1}$$

In terms of the matching procedure, the spacing between the slugs adjusts the magnitude that can be matched, while the line in front of the load adjusts the phase.

Slug matching is also used in nonlinear device (and circuit) characterization, usually to build empirical device models. In the load-pull and source-pull approach, air coaxial tuners are used to provide as large of a range of impedances to a device under varying input power and DC conditions. In this case, the slugs are actually metal pieces that move coaxially and slide along the z-axis. Fig. 5 shows a photo of a Focus Microwave fundamental frequency tuner, with a detail of the slug geometry.

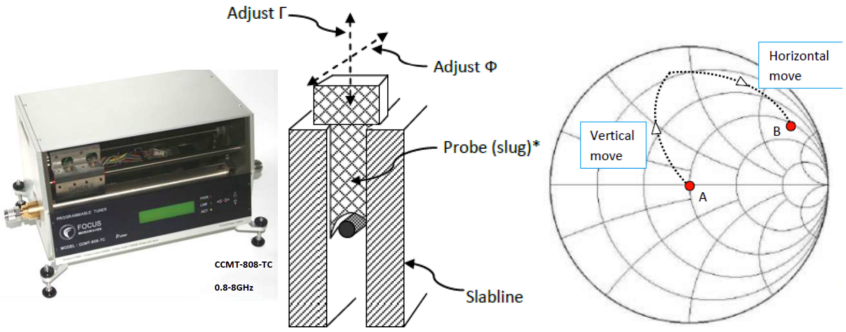


Fig. 5 Mechanical load-pull tuner with slug geometry and function shown on a Smith chart.

1.5 Single-Section Line Matching

If an impedance lies in a certain area of the Smith chart, it is possible to do a single-line matching section. In this case both the impedance and the electrical length need to be determined, and this type of match is relatively broadband. First, we need to decide what the possible range of impedances that are available could be. For example, in microstrip, it is usually about 20-90 Ω. Fig. 6 shows the range of impedances that can be matched with a single section of line -– they lie in the $r = 1$ and $g = 1$ circles (why?).

A possible method for determining the required characteristic impedance and line length is sketched in Fig. 6. First, r_1 is determined using a bisector of the line between the normalized load impedance and the center of the chart. That determines Z_T. Now we need to renormalize Z to Z_T, which gives Z'. Moving to the real axis by θ gives r_2 which now renormalized to 50 Ω gives the desired line Z_0. In your homework, you will go over a different approach using the simulator.

1.6 Broadband Matching

The general idea in making a broad-band matching circuit is to transform the impedance in geometric steps. Typically the bandwidth is proportional to the number of steps. The transformation can be done with lumped elements, quarter-wave

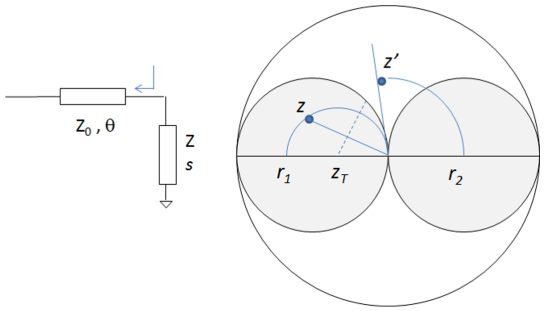


Fig. 6 Matching using a single section of transmission line. The Smith chart shows a possible method of determining the required Z_0 and θ in order to match a normalized impedance $z = Z/Z_0$.

sections or stub networks. On standard textbook method is to use an exponentially-tapered nonuniform line, or a Klopfenstein taper (see Pozar's discussion for more information). This method can only be used to match purely real impedances. Another method that can be used to transform real impedances are coaxial transformers, the simplest is illustrated in Fig. 7 (left). At the input, the outer conductor of the top transmission line is connected to the inner conductor of the bottom line, and the two inner conductors are connected at the output. This results in the total voltage at the input being the sum of the two line input voltages (in series), and the total current at the output is the sum of the two inner conductor currents (in parallel). Thus, the voltage and current transformation ratios are equal to 2 and the impedance transformation ratio is 4:1. This means that such a transformer matches a 12.5-Ω load to a 50-Ω line.

The interesting thing is that this is true over a very large bandwidth (often several decades) if implemented in coaxial line. Notice that there is nothing specified for the electrical length θ, except that the two lines are exactly the same length, giving the same delay. For very short lines, they are a lumped element effectively, and the coaxial lines are in that case wound around a magnetic core which is used to suppress unbalanced currents in the shield. At the higher frequency end, the line length is not the limitation, but rather the parasitics of the connections at the input and output, which will effectively unbalance the delays. The line impedance is $Z_0/\sqrt{N}$, so for a 4:1 transformer as in the figure on the left, $Z = 25\,\Omega$.

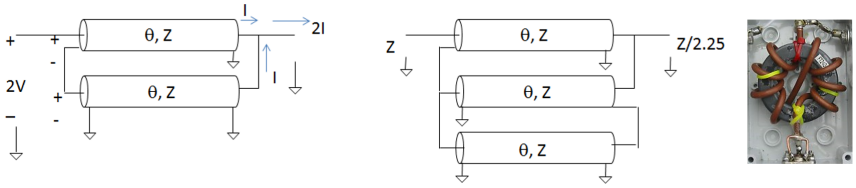


Fig. 7 Coaxial 4:1 impedance transformers. Left: 4:1 impedance transformer with Z=25 Ω, Middle: 2.25:1 with Z=33.3 Ω. Right: photo of a 3-30 MHz transformer balun.

Other transformation ratios are also possible, but it turns out that $\sqrt{N}$ needs to be a rational number. For example, 2:1 is not possible, because $\sqrt{2}$ is not a rational number, but 2.25 is possible and it is close enough. The design of a 2.25:1 transformer is shown in Fig. 7 in the middle. A reference for these transformers is given at the end of this lecture. These types of transformers can also be implemented in non-coaxial media, usually at the expense of bandwidth and with some grounding issues that need to be solved.

2 Useful 3-Port and 4-Port Networks for Active Circuits

Above, we reviewed matching circuits, which are 2-port networks. They are passive and can be reciprocal and lossless, but are generally not matched (otherwise you would not need them). As you have studied in previous classes, a 3-port matched, lossless and reciprocal circuit is not possible, but useful 3-port networks include unmatched Tee dividers (e.g. for antenna feeds), lossy Wilkinson dividers, and non-reciprocal circulators. Another useful 3-port network is a bias-tee, which we will need to provide DC power to a transistor so that it can either produce RF power (in an oscillator) or amplify RF power at the expense of the DC input (in any amplifier).

2.1 Bias Networks

The active device in an amplifier, oscillator, mixer, etc. need to be connected to one or more bias supplies, which ideally not affect the RF performance of the circuit. Designing good biasing circuits is a large part of amplifier design, and the following are critical design parameters:

1) the bias network needs to be "invisible" to the RF waves, i.e. as close to an open circuit as possible. The reason is that we cannot afford any of the RF power to be lost in the biasing circuit and power supply;
2) the DC bias needs to be isolated from the RF circuit, i.e. we do not want the DC voltage to be present at the RF input (e.g. we might be dealing with a 2-stage amplifier, and the previous stage requires a different voltage);
3) finally, the DC bias circuit should be designed over the entire frequency range where the device has gain, so as not to cause instabilities.

In order to satisfy the first criterion, the dc bias lines could consist of an inductor with a value chosen to present a high impedance at the RF, as shown in Fig. 8. It is difficult to make a high-valued inductor at microwave frequencies due to parasitic capacitance. Another option is that the bias lines have the characteristics of a low-pass filter (review basic low-pass filters if needed).

In order to satisfy the second requirement, a dc blocking capacitor needs to be added to the circuit, and needs to be taken into account in the design. Capacitors are not ideal shorts at microwave frequencies (they have parasitic series inductance and shunt conductance). The bias circuit can be integrated with the amplifier matching circuit, or alternatively, an external biasing circuitry can be used. External bias

circuits are often referred to as Bias Tees. These devices are expensive if they cover a broad bandwidth, and usually have current limitations. The reason is the inductor in the DC path, which needs to be made of thin wire so as not to have appreciable parasitic capacitance.

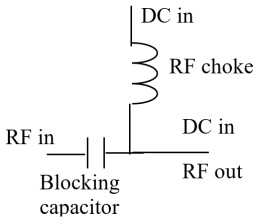


Fig. 8 A bias-Tee ideal equivalent circuit.

Commercial bias Tees are fairly large and have typically SMA connectors at the two RF ports. Often, biasing circuits are part of amplifier design, and some examples are shown in Fig. 9.

The dc biasing circuit should be taken into account when analyzing stability, i.e. it is part of the input and output network. Even though it is designed to present a high impedance to the RF signal at the design frequency (convince yourself why this is so), it is not a real open circuit. For example, at a frequency other than the design frequency, the quarter-wave shorted line is not a quarter-wavelength long, and therefore is not an open circuit to the RF signal. There is also some loss in the blocking capacitor – the DC blocking capacitor has lead inductance and some resistance, and it will not be perfectly matched to the input RF 50-Ω line. In the grounded capacitor implementation (righthand side of Fig. 9), the capacitor and via hole have inductance that is usually not well characterized, and this also determines the quality of the open circuit presented to the RF signal.

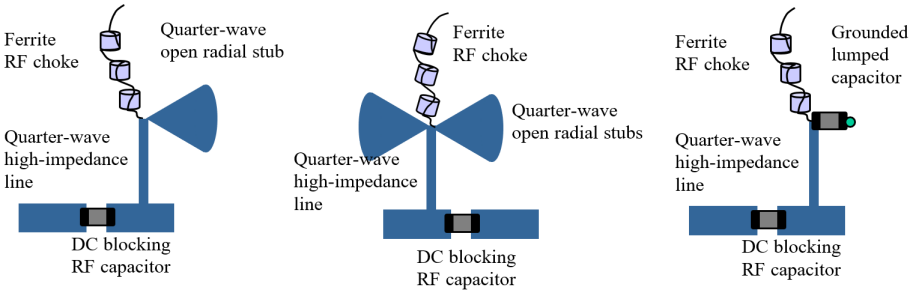


Fig. 9 Examples of microstrip biasing circuits.

The ferrite choke is an inductor at lower frequencies and is effective at choking frequencies up to a few hundred MHz. This is important, since the bias lines can be good antennas for broadcast, WiFi etc. signals. At microwave frequencies, however, the ferrite is just a large resistor (the material is very lossy), so the RF currents will be very attenuated and will not reflect back into the circuit. However, the power is lost and any power flow into the ferrite lines should be minimized.

A difficult problem is a broadband bias line. In principle, a good inductor with several hundred nH inductance would solve the problem, but microwave inductors typically do not work above a few GHz due to parasitic capacitance. If loss is not an issue, however, the Q factor of the inductor can be reduced by adding resistors or ferrites and very broadband bias networks can be made. A very broadband amplifier from Agilent (0-40GHz) needs very broadband bias lines, as shown in Fig. 10. The tiny cone-shaped inductors with ferrite loading are from Piconics, but Coilcraft makes them as well. The idea is that the Q is greatly reduced at high frequencies, so the resonance due to the parasitic capacitance is not relevant.

These broadband inductors are conical, so effectively they look like a continuous set of series inductors with progressively higher inductance values and, correspondingly, progressively lower self-resonance frequencies. Effectively, this is a distributed series of bandpass filters, shown in Fig. 10. You can make a broadband bias line using several series inductors in this fashion.

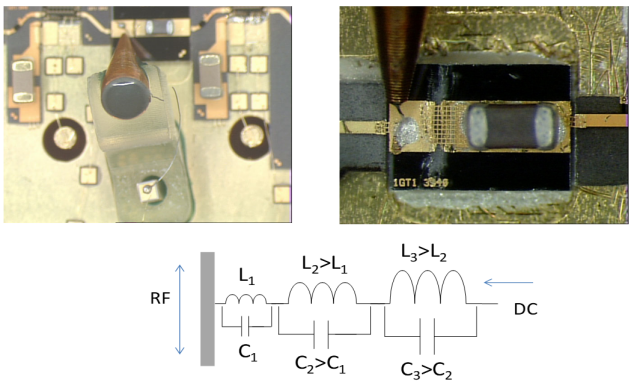


Fig. 10 Top: Photographs of miniature conical coil with ferrite loading (www.piconics.com). Bottom: discrete series inductors which in the continuous limit behave like the conical inductor. If a ferrite is added, it increases inductance at the lower frequencies and adds loss at the higher frequencies.

2.2 Directional Couplers

4-port networks can be made lossless, reciprocal and matched, and there are several useful examples, such as various 90-degree and 180-degree couplers. Consider a 4-port network that is matched, reciprocal and lossless, with two-plane symmetry

(such as the coupled-line coupler in Fig. 11 where the symmetry planes are shown in dashed line). The scattering matrix can be written in the following form:

$$[S] = \begin{bmatrix} 0 & \alpha & \beta & \gamma \\ \alpha & 0 & \gamma & \beta \\ \beta & \gamma & 0 & \alpha \\ \gamma & \beta & \alpha & 0 \end{bmatrix} \tag{2}$$

where we have used the reciprocal condition and symmetry. Next, we can use the lossless condition for the inner products of the various columns of the S matrix, which gives $\beta^*\gamma = \gamma^*\beta$, which in turn implies that $2\Re e\{\beta^*\gamma\} = 0$. This expression can be expanded to:

$$2\Re e\{(x_\beta - jy_\beta)(x_\gamma + jy_\gamma)\} = x_\beta x_\gamma + y_\beta y_\gamma = 0 \tag{3}$$

where $x_{\beta,\gamma}$ and $y_{\beta,\gamma}$ are the real and imaginary parts of β, γ, respectively. By looking at this expression, which is that of an inner product, we can conclude that β and γ are orthogonal, or 90 degrees out of phase, or alternatively that one of them is zero. The same analysis for the lossless condition of other columns of S result in α and β, as well as α and γ being orthogonal. Since it is impossible for 3 two-dimensional vectors to be mutually orthogonal, we conclude that one of them has to be zero. If $\gamma = 0$, this means that port 4 is isolated. Furthermore, the other two ports are in quadrature. Since the lossless condition also gives

$$|\alpha|^2 + |\beta|^2 = 1 \tag{4}$$

which implies that the power is divided between ports 2 and 3 (not necessarily equally).

2.3 Branch-Line Hybrid Coupler

A branch-line coupler is shown in Fig. 13, and the usual odd and even mode analysis can be used to find the S-parameters. However, if we know that port 2 is isolated, it means there is no current flowing into it, and therefore we can short this port. Later, we will need to prove that the solution is consistent with there being no current in the short.

First, by shorting port 2, we find the equivalent circuit on the right in Fig. 11. The load at port 3 is $Z_0/2$, so the impedance at port 1 after the quarter-wave section is $(Z_0/\sqrt{2})^2/(Z_0/2) = Z_0$. Therefore, port 1 is matched. Since port 3 sees two Z_0 loads in parallel, the current through them is the same, so the power divides equally. Therefore, this is a 3-dB coupler. Further, since the impedance of the quarter-wave section of line between ports 3 and 4 is Z_0, the voltage at port 4 is the same in magnitude as that at port 3, but 90° out of phase. So, the coupler is a 90-degree 3-dB coupler.

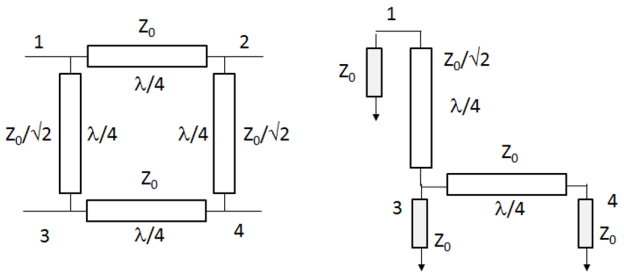


Fig. 11 Branch-line coupler, port 2 is isolated, and ports 3 and 4 are in quadrature. On the right is the equivalent circuit when port 2 is short-circuited.

The current at the input of port 3 lags the voltage at port 1 by 90° so $I = -jV_1/(Z_0/\sqrt{2})$. Since the current divides equally at port 3, $V_3 = IZ_0/2 = -jV_1/2$. Therefore, voltage at port 3 lags voltage at the input by 90°. This happens because the impedance that the current I sees is real. Now we need to show that the current in port 2 is indeed zero, and therefore that removing the short does not affect the circuit. We can compare V_1 and V_4 as follows: $|V_1| = \sqrt{2}|V_4|$ since half of the power goes into port 4. Also, $\angle V_1 = \angle V_4 - 180°$ because of two quarter-wave sections. Since the impedance of the line between ports 2 and 4 is lower than the impedance between 1 and 4 by a factor of $\sqrt{2}$, this compensates for the lower voltage at port 4, and the two current magnitudes at port 2 are equal but of opposite sign. Therefore, no current is flowing into the short at port 2. Can you apply a similar reasoning to quickly solve the rat-race coupler circuit?

2.4 Coupled Microstrip Lines

Coupled lines are parallel transmission lines that are close together, so that the electric and magnetic fields on the lines are not independent and couple from one line to the other. They are useful as components in broadband directional couplers (both 90° and 180°) and bandpass filters. The bandwidth of coupled line couplers can be much larger (decades) than in the case of a branch line or rat-race (20-30%).

Figure 12 shows electric and magnetic field lines for two coupled microstrip lines. If the left line is connected to a voltage source, there will be electric coupling to the right line through electrostatic induction. If current is allowed to flow on the left line, there will be magnetic coupling between the two lines through Faraday's law. These two induction mechanisms are responsible for coupled-line behavior at all frequencies. The second line can also be connected to a voltage source, and there are two possible situations, shown in the figure and called the"odd" and "even" mode, referring to the symmetry of the electric and magnetic fields.

Now let us go through a little mental experiment. Assume you have two lines, and you quickly bring equal length parts of them close together, as shown in Fig. 13. If a

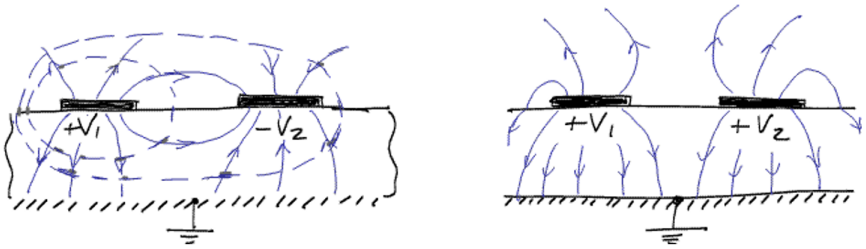


Fig. 12 Sketch of coupled microstrip lines showing the electric and magnetic fields for the odd and even modes.

wave is incident at port 1, you can imagine it coupling capacitively to port 3 as soon as the wave reaches the coupled section. Then a wave is established on the coupled section and propagates until the line splits again, with some phase delay along the coupled part. It is not obvious if there will be some reflected wave at the input port 1. By adding a second line close to line 1, we have changed the impedance of that part of line 1, so one would expect a reflection. In fact, line 1 looks thicker in the coupled part, so it seems as if the impedance went down and there will be a reflection. If we want to eliminate the reflection, we can narrow both lines in the coupled region until we match port 1. This will raise the inductance and lower the capacitance of the lines in the coupled region.

Note that optical fiber couplers look similar to these, but operate quite differently. In the case of fiber couplers, the evanescent mode (exponentially decaying away from the core of the fiber) couple, and the amount of coupling depends on the length of the line and the isolated port is port 3. In our case, the isolated port is port 4.

The fact that the circuit in Fig. 13 has a port isolated and the other two 90-degrees out of phase is independent on the length of the coupled section. Port 1 is matched, port 4 isolated and ports 2 and 3 in quadrature, and this will not change with the length of the coupled line section length. The only thing that changes is the coupling coefficient. Note that this is quite different than in the case of the branch-line coupler, which depends on phase-dependent additions and cancellations at the ports, resulting in frequency dependence.

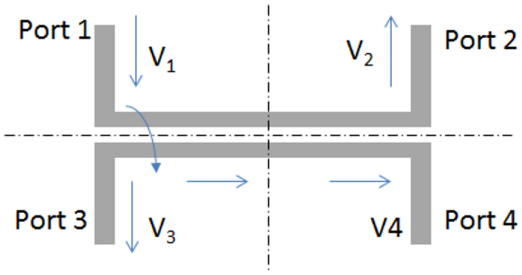


Fig. 13 Voltage waves for a coupled-line coupler.

2.4.1 Quasi-static (long wavelength) analysis

Consider a short edge coupler, with all ports properly terminated, and an incident wave V_1 at port 1, as in Fig. 14 We will make a few assumptions:

- Wavelength is much longer than the lines;
- Coupling is weak, which means that the coupled line 2 does not affect the voltage and current on line 1; and
- Because the coupler is short, V_1 and I_1 are the same all along line 1.

Capacitive coupling

The voltage on the line 1 will raise the voltage on line 2, so current will flow towards both of the terminating resistors. Since there is a capacitance between lines 1 and 2, the currents in line 2 will lead V_1 by 90°, which means that the current will flow in the $+z$ direction at port 4, and the $-z$ direction at port 3. Now we can write a transmission-line equation for the current on the second strip:

$$\frac{\mathrm{d}I_2}{\mathrm{d}z} = \mathrm{d}I_2' = -j\omega C_m V_1, \tag{5}$$

where C_m is the distributed mutual capacitance, and the prime refers to differentiation with respect to z. By symmetry, there is no current in the middle of the second strip, and this gives a reference point that can be useful to integrate the equation which we assume is linear in z since the wavelength is long compared to the coupled line length:

$$I_{C2}(z) = -j\omega C_m V_1 z. \tag{6}$$

The current along the strip varies linearly with distance, because we can think of it as physically coming from a row of small distributed capacitors between the strips. More current is accumulated as we approach the end of the line. The mutual capacitance has to be negative in order for the current to have the right phase. The current magnitude at the end is:

$$|I_{C2}(\pm\ell/2) = \omega C_m V_1 \ell/2. \tag{7}$$

Inductive coupling

The lines are a weakly-coupled transformer. When a current flows in line 1, there is a magnetic field around it and flux between the lines, resulting in mutual inductance. This causes, through Faraday's law, a circulating current in line 2 and the terminations. The current will oppose the flux. There is a voltage now across the resistor in port 3, and an equal but opposite voltage at port 4. By symmetry, the voltage in the middle is zero, and again this is the reference point for integration of the transmission-line equation:

$$\frac{\mathrm{d}V_2}{\mathrm{d}z} = \mathrm{d}V_2' = j\omega L_m I_1, \qquad \textit{and} \qquad V_{L2}(z) = j\omega L_m I_1 z. \tag{8}$$

where L_m is the distributed mutual inductance. The inductively induced voltage should lag I_1, and this means that the mutual inductance is positive. The voltage magnitude at the termination is:

$$|V_{L2}(\pm\ell/2) = \omega L_m I_1 \ell/2. \tag{9}$$

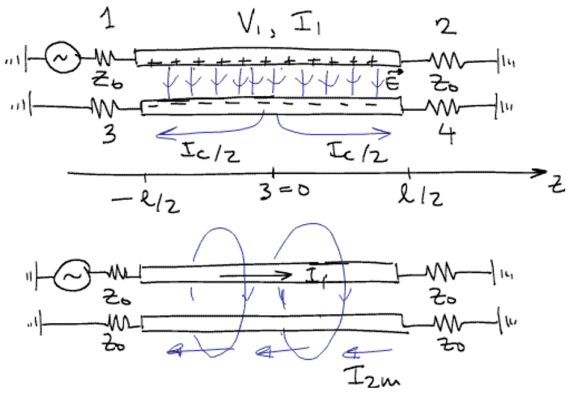


Fig. 14 Voltage waves for a coupled-line coupler.

The capacitively and inductively coupled voltages at port 3 are in phase, but at port 4 they are out of phase, and tend to cancel. There is complete cancellation under the following condition between the terminating impedance and the mutual capacitance and inductance (keeping in mind that the mutual capacitance is negative):

$$\frac{L_m I_1}{-C_m V_1} = Z_0 \tag{10}$$

For a short coupler, $V_1/I_1 = Z_0$, resulting in

$$-\frac{L_m}{C_m} = Z_0^2 \tag{11}$$

This means that port 4 is isolated if we design the coupled lines to have this type of mutual inductance and capacitance. Notice that the result does not depend on the length of the line under these assumptions.

2.4.2 Transmission-line (distributed, high-frequency) coupled-line coupler analysis

What types of waves propagate on a line that is not very short compared to a wavelength? We can write the transmission line equations for the lines, but with two lines, there will be two voltages, two currents, and distributed mutual and self-

capacitances and inductances. We can write the transmission line equations in vector form as:

$$\underline{I'} = -j\omega \underline{\underline{C}} \cdot \underline{V} \qquad \text{and} \qquad \underline{V'} = -j\omega \underline{\underline{L}} \cdot \underline{I} \tag{12}$$

where the $\underline{V}, \underline{I}, \underline{\underline{C}}, \underline{\underline{L}}$ are given by:

$$\underline{V} = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}, \quad \underline{I} = \begin{bmatrix} I_1 \\ I_2 \end{bmatrix}, \quad \underline{\underline{C}} = \begin{bmatrix} C_s & C_m \\ C_m & C_s \end{bmatrix}, \quad \underline{\underline{L}} = \begin{bmatrix} L_s & L_m \\ L_m & L_s \end{bmatrix}. \tag{13}$$

The subscript "s" stands for "self" and "m" for "mutual". The current can be eliminated, just as in the single line case, to give a vector wave differential equation:

$$\underline{V''} = \omega^2 \underline{\underline{C}} \cdot \underline{\underline{L}} \cdot \underline{V} \qquad \text{where} \qquad \beta^2 \underline{V} = \omega^2 \underline{\underline{C}} \cdot \underline{\underline{L}} \cdot \underline{V} \tag{14}$$

is the characteristic equation. The solutions are, just in the case of a single line, or the form $V\exp(\pm j\beta z)$. The eigenvalues of this equation give the propagation constants, and the eigenvectors are the voltage modes. We have assumed that $\underline{\underline{C}}$ and $\underline{\underline{L}}$ are symmetric matrices, so their product is also a symmetric matrix. Going back to your linear algebra class, you can conclude that the eigenvalues are real and the eigenvectors orthogonal.

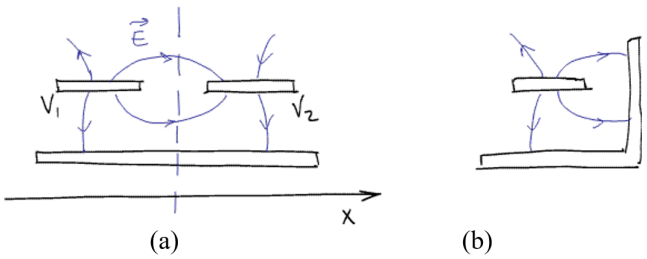


Fig. 15 Odd mode field lines (a). Equivalent boundary conditions for odd mode (b).

Even and odd modes are orthogonal, and this can also be concluded from physical reasoning. Assume that the two eigenvalues are different (propagation constants not the same). This is true in microstrip, but not in coax. Now let us invert the x-axis in Fig. 15. This is the same as replacing index 1 by index 2 and vice versa. By symmetry, the line has not changed, so this new pair of voltages must also be a solution. In fact, it has to be the same as before because the propagation constants stay the same. This means the mode does not change when the indices are interchanged. This in turn is only possible for odd and even voltages, i.e. $V_1 = V_2$ and $V_1 = -V_2$. If the mode is odd, there is a sign change, but this does not change the mode, since the modes are

defined only to an arbitrary scalar constant. Since the eigenvalues are odd and even, this means that the voltage and current eigenvectors will be odd and even as well, and written as follows:

$$\underline{V_e} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{\mp j\beta_e z} \qquad \underline{V_o} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{\mp j\beta_o z} \tag{15}$$

$$\underline{I_e} = \pm \underline{V_e}/Z_e \qquad \underline{I_o} = \pm \underline{V_o}/Z_o \tag{16}$$

$$\beta_e = \sqrt{(L_s + L_m)(C_s + C_m)} \qquad \beta_o = \sqrt{(L_s - L_m)(C_s - C_m)} \tag{17}$$

$$Z_e = \sqrt{\frac{L_s + L_m}{C_s + C_m}} \qquad Z_o = \sqrt{\frac{L_s - L_m}{C_s - C_m}} \tag{18}$$

The Ls and Cs can be solved using quasi-static analysis, which we will not do. However, consider Fig. 15(b) which shows the equivalent boundary conditions for the odd mode. There is no y-oriented E-field component in the symmetry plane, so a metal wall can be placed there. This results in increased capacitance per unit length compared to a single line. This means that the mutual capacitance has to be negative, referring to the expressions above. Likewise, inductance per unit length decreases, so the mutual inductance is positive.

Analyzing a coupler in terms of odd and even modes makes things simpler, since we can write the voltages and currents on the uncoupled lines in terms of the even and odd modes as well and in this case the even and odd mode impedances are the same and equal to the characteristic impedance of the transmission line. This suggests splitting the circuit into two independent problems: one for the odd mode and one for the even mode, and treat them separately as long as nothing perturbs the symmetry.

The simplest couplers are quarter-wave long at the design frequency. Assume odd and even mode propagation constants are the same (so that both are quarter-wave long), you can then easily calculate the odd and even mode reflection coefficients from two quarter-wave transformers, one with Z_e and the other with Z_o:

$$\rho_{e,o} = \frac{z_{e,o}^2 - 1}{z_{e,o}^2 + 1} \qquad \textit{and} \qquad \tau_{e,o} = \frac{-2jz_{e,o}^2}{z_{e,o}^2 + 1} \tag{19}$$

where $z_{e,o} = Z_{e,o}/Z_0$ are the normalized odd and even mode impedances. The above expressions combine to give the two-port coupled-line coupler S-parameters as $s_{11} = (\rho_o + \rho_e)/2$, etc. for the other parameters.

3 Microwave Transistors

The most commonly used active devices made in III-V semiconductors at microwave frequencies are field effect transistors (FETs). In the past, the workhorse was the Metal Semiconductor Field Effect Transistor (MESFET), while currently the dominant one is a variation, called a High-Electron Mobility Transistor (HEMT). The GaAs FET device physical cross section is shown in Fig. 16(a), with a typical electrode layout in Fig. 16(b). The MESFET/HEMT is a unipolar device, which means that there is only one type of carrier. The device has three terminals: the source, gate and drain. The gate length is the length that the electrons need to travel between the drain and the source and is usually a fraction of a micrometer. For example, in the WIN GaAs PQG3 process that we will use in the designs for this class, the gate length is 150 nm and determines the highest operating frequency. The gate width is in the direction out of the paper referring to figure and can be hundreds of micrometers long, which is called the "gate width". It determines the current that can flow through the device; larger gate widths (longer gate fingers), and more gate fingers in parallel increases the current and therefore output power.

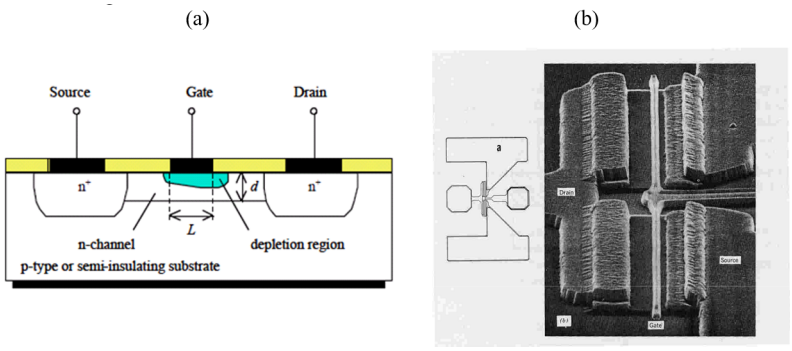


Fig. 16 A cross section of a microwave FET, specifically a MESFET: (a) and photograph and electrode layout (b).

The three electrodes are deposited on an n-type GaAs epitaxial layer which is grown on a semi-insulating substrate. The epitaxial layer is on the order of 0.1 μm thick and the doping is $10^{16} - 10^{17}$ cm^{-3}. The source and drain are ohmic contacts (low resistance, usually made of a gold-germanium alloy), and the gate is a Schottky contact. Associated with the Schottky barrier is a depletion region which affects the thickness of the conducting channel. In a so called depletion mode (or D-mode) device, the gate is biased negatively with respect to the source, and the drain positively. When the voltage is changed on the gate, the thickness of the channel changes, and this controls the current between the drain and the source. The device is normally on, and applying a negative gate voltage pinches the channel, depleting

the carriers and turning the device off. There are also enhancement mode (E-mode) devices, which are normally off, so applying a gate voltage turns the device on.

3.1 MESFET/HEMT Equivalent Circuit Models

When you buy a MESFET or HEMT, it can come in a package or in chip (die) form. You will also get measured *S*-parameters at a few different bias points for a certain frequency range. These *S*-parameters are measured usually with the source terminal grounded and the drain and gate looking into 50 Ω, so they are two-port parameters. The $|s_{21}|$ parameter corresponds to the gain of the device in common-source configuration. The amplitude and the phase of all four parameters are given at many discrete frequencies. Another way to represent the transistor is with an equivalent circuit, like you have probably done in your circuits classes. The idea behind equivalent circuits is to model the device over a range of frequencies with invariant parameters. Let us begin with the simplest linear (small-signal) equivalent circuit, for which it is simpler to use admittance parameters, given by **I=YV**.

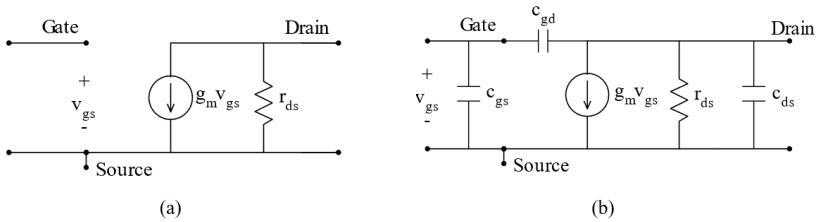


Fig. 17 (a) Low-frequency MESFET model. (b) High-frequency approximate MESFET equivalent circuit. These circuits are called the intrinsic equivalent circuits because additional parasitics from the package are not included.

The *Y*-parameters may be converted to *S*-parameters using the formulas from e.g. Pozar's or Collin's books (or you can derive them). At very low frequencies, say below 1 GHz, the capacitances and inductances associated with the MESFET/HEMT are quite small, and we can assume they are negligible. The same is true for the resistive losses. The result is a simple low-frequency model shown in Fig. 17(a). This model has an infinite input impedance and cannot be matched at the input. What is the order of magnitude of the elements of this circuit?

Let us look at a few examples. The Qorvo TGF2960 device is a 0.5 W GaAs HFET with *S*-parameters at the lowest 100 MHz frequency with V_{DS} = 8 V, I_{DS} = 100 mA given by:

$$\begin{aligned} s_{11} &= 0.677\,\text{dB}\angle -7.9^\circ \approx 1\angle 0^\circ \\ s_{21} &= 20\,\text{dB}\angle -175.75^\circ \approx -10 \\ s_{12} &= -42\,\text{dB}\angle 72.5^\circ \approx 0 \\ s_{22} &= -5.6\,\text{dB}\angle -19.43^\circ \approx 0.5\angle 0^\circ \end{aligned} \tag{20}$$

If we wish to find the values of the elements in the equivalent circuit, we would first solve for the S-parameters of the equivalent circuit in terms of the unknown elements, and then set the expressions equal to the known S-parameters, thus getting a system of equations. Finding the expressions for the S-parameters of the equivalent circuit can be quite complicated, and usually the Y-parameters are found and then converted to S-parameters. For the low-frequency model from Fig 17 (with no capacitors), the Y-parameters are:

$$\underline{\underline{Y}} = \begin{bmatrix} 0 & 0 \\ g_m & g_{ds} \end{bmatrix} \tag{21}$$

and the S-parameters are obtained as follows:

$$\underline{\underline{S}} = \begin{bmatrix} 1 & 0 \\ \frac{-2g_m}{1+g_{ds}} & \frac{1-g_{ds}}{1+g_{ds}} \end{bmatrix} \tag{22}$$

Since the specification sheets provide the S-parameters, one can find the conductance values; note that they are normalized to $1/50\,\Omega = 0.02\,\mathrm{S}$.

Next we will compare an Avago (now Broadcom) MESFET chip with reasonably standard parameters, that at 500MHz has S-parameters at $V_{DS} = 3\,\mathrm{V}$, $I_D = 20\,\mathrm{mA}$ as follows:

$$\begin{aligned} s_{11} &= 0.97\angle 20^\circ \approx 1 \\ s_{21} &= 5\angle -166^\circ \approx -5 \\ s_{12} &= 0.029\angle 70^\circ \approx 0 \\ s_{22} &= 0.52\angle -110^\circ \approx -0.5 \end{aligned} \tag{23}$$

Notice that some manufacturers provide magnitude in dB, and some do not, so it is good to pay attention. Finally, consider the Qorvo (RFMD) FPD HEMT device, which at 5 V and 300 mA has the lowest available frequency of 50 MHz S-parameter values as follows:

$$\begin{aligned} s_{11} &= 0.946\angle -24^\circ \approx 1 \\ s_{21} &= 36\angle -159^\circ \approx -36 \\ s_{12} &= 0.006\angle 79^\circ \approx 0 \\ s_{22} &= 0.175\angle -145^\circ \approx -0.175 \end{aligned} \tag{24}$$

Going through the same approximations as for the FETs above, we get the parameters calculated from the low frequency measured data to be $g_m = 0.875\,\mathrm{S}$, $G_{ds} = 24.5\,\mathrm{mS}$ and the voltage gain $A_v = 30$. If you look at the measured data at $f = 600\,\mathrm{MHz}$ for the same bias point, however, these are the values:

$$\begin{aligned} s_{11} &= 0.821\angle -144^\circ \\ s_{21} &= 10.59\angle 98.5^\circ \\ s_{12} &= 0.032\angle 38^\circ \\ s_{22} &= 0.568\angle 174^\circ \end{aligned} \tag{25}$$

Referring back to the discussion about the low-frequency transistor model, you can see that the simple equivalent circuit from before cannot be used. There are capacitances that are already quite pronounced at 600 MHz. The reason is that the MESFET device has gain up to much higher frequencies (12 GHz) and can give at most a hundred mW of power, while the RFMD HEMT is a 1 W device for lower frequency operation.

The most basic high-frequency intrinsic equivalent circuit model is shown in Fig. 17(b). The same method can be used to calculate the circuit parameters from measured S-parameters. You will calculate the Y matrix for this circuit in Project 2. Some important additional parameters are the gate, source and drain contact resistances, and these are measured usually at DC, though their values will change at high frequency somewhat due to the skin effect. When the gate, drain and source+drain are shorted, resistances R_a, R_b and R_c are measured, respectively. The contact resistances can then be found from:

$$R_G = R_c - \sqrt{R_c^2 - R_C(R_a + R_b) + R_a R_b}, \quad R_D = R_b - R_G, \quad R_S = R_a - R_G \tag{26}$$

The depletion capacitance of the Schottky barrier gate is represented by C_{gs} and C_{gd}. Usually C_{gs} is much larger. The reason is that the positive voltage on the drain causes the depletion region on the drain side to be wider than on the source side. Also, the separation between the drain and gate contacts is usually about 1 μm larger than that between the source and gate. The capacitance between the source and drain is primarily through the substrate, and is not negligible because of the high dielectric constant of GaAs of 13. The resistance of the gate is significant because the gate contact is long and thin, and a typical value is $10 - 15\,\Omega$. The usual figure of merit for the transistor is the voltage gain $A_V = g_m/g_{ds}$. Since both conductances are proportional to the gate width, the voltage gain does not depend on the width. This is important in MMICs, where there is complete control over gate widths, but gate lengths are fixed by the fabrication process. The gate length determines the maximum operating frequency of the device (directly, the RC time constant). An experimentally obtained formula is:

$$f_{max} = \frac{33 \cdot 10^3}{L}\ \text{Hz}, \tag{27}$$

where L is the gate length in meters. Several cutoff frequencies are commonly used. f_T is the cutoff frequency when the short-circuited current gain of the device drops to unity. This parameter is often used, but never measured, since a microwave transistor tends to oscillate with a short-circuit load. If the input current for the high-frequency equivalent circuit is i_{in}, then we can write:

$$i_{in} \approx j\omega C_{gs} \cdot v_{gs} \qquad \text{and} \qquad i_{out} \approx i_{ds} \approx g_m \cdot v_{gs} \tag{28}$$

$$|i_{out}/i_{in}| = 1 \quad \rightarrow \quad \omega = \omega_T = \frac{g_m}{C_{gs}} \quad \text{and} \quad f_T = \frac{g_m}{2\pi C_{gs}} \tag{29}$$

The two most important parameters for the high-frequency performance are therefore g_m and C_{gs} −− large g_m and small C_{gs} result in a high cut-off frequency. A typical procedure used to calculate the cutoff frequency is to derive the short-circuit gain from the measured S-parameters, and extrapolate this curve to the value of the gain equal to 0 dB, illustrated in Fig. 18. This gives a simplified 6 dB/octave response, although the actual one is obviously more complicated (we used only approximate formulas).

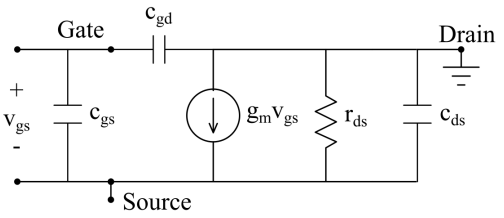


Fig. 18 High-frequency approximate MESFET equivalent circuit with shorted drain to determine the f_T.

The maximum frequency of operation is higher than the cutoff frequency, and is defined as a frequency for which a negative resistance (oscillation) can be produced. The two frequencies are related by

$$f_{max} = \frac{f_T}{2\sqrt{r_1 + f_T r_2}}, \tag{30}$$

where $r_1 = (R_G + R_S + R_i)/R_{ds}$, $r_2 = 2\pi R_G C_{gd}$ and the different resistances are those of the gate and source contacts, R_i the intrinsic resistance between source and gate, and R_{ds} the drain-to-source resistance.

The unilateral transistor gain as a function of frequency can be expressed in terms of the cutoff frequency as:

$$G_u \approx \left(\frac{f_{max}}{f}\right)^2. \tag{31}$$

In this approximation, the gain is 1 when $f = fmax$. The maximum frequency is usually two to three times higher than the cutoff frequency. In order to obtain a high f_{max}, the cutoff frequency needs to be maximized, as well as the ratio of channel resistance to $(R_G + R_S + R_i)$, and C_{gd} needs to be minimized. The transit time is decreased by decreasing the gate length L, but this also results in a decrease in channel depth in order to maintain a geometry that gives a high g_m. In turn, this means that the doping in the channel must increase, to maintain a low channel resistance. The limit on the doping is set by the avalanche breakdown in the gate-drain region which has the highest fields, and it is about $5 \cdot 10^{17}\,\mathrm{cm}^{-3}$. This discussion clarifies why it is difficult to make high-frequency power devices.

A method for reducing the series resistance in the source is to recess the gate, by making a mushroom type structure for the gate metallization, by self-alignment. The cutoff frequency can be directly related to the transit time of electrons under the gate by the following approximate argument. Assume a small positive change in the gate voltage Δv_g. This results in: (1) the gate charges up by $\Delta q = C_{gs} \cdot \Delta v_g$ since it is one of the capacitor electrodes; (2) the other electrode of the capacitor is the channel, so the same amount of negative charge must be drawn into the channel. Negative charge in the channel means an increase in carrier (electron) density, so (3) the current through the channel (i_{ds}) increases. The time that the electrons take to transit the gate region is found from $i_{ds} = \Delta q/\tau_T = G_{gs} \cdot \Delta v_g/\tau_T$, and from the definition of transconductance, the following can be written:

$$\frac{g_m}{C_{gs}} = \frac{1}{\tau_T}, \tag{32}$$

where the left-hand side is the significant ratio for the cutoff frequency that we derived earlier. We can therefore express the cutoff frequency in terms of the transit time as follows:

$$f_T = \frac{g_m}{2\pi C_{gs}} = \frac{1}{2\pi\tau_T} \approx \frac{v_{sat}}{2\pi L}, \tag{33}$$

where v_{sat} is the carrier saturation velocity. So, this gives a very simple rule: if you wish to make a device with a high cutoff frequency, you need to increase the saturation velocity and decrease the gate length. The saturation velocity in bulk GaAs is limited to about 10^7 cm/s, and to overcome that the semiconductor material under the gate must be modified. This is done is High Electron Mobility Transistors (HEMTs). If you wish to make a device with a high cutoff frequency, you need to:

- increase the saturation velocity of electrons in the channel and
- decrease the gate length of the device.

It is somewhat obvious what issues need to be solved in decreasing the gate length and that they are purely technological, i.e. require better photolithography. To increase the saturation carrier velocity, however, requires a new device design. What is the limit in saturation velocity in a GaAs FET?

If no collisions occur, electrons in GaAs are accelerated by the electric field and follow Newton's second law, with the mass replaced by the effective mass m^*:

$$F = eE = m^* \frac{\mathrm{d}v}{\mathrm{d}t}, \quad \rightarrow \quad v = \frac{eE}{m^*} \cdot t, \tag{34}$$

where e is the electron charge. At room temperature, the mean free path for the electrons can be estimated from the measured mobility, and the value of about $0.1\,\mu$m is a reasonable value. If the electron starts at $x = 0$, then at time τ it has traversed the entire gate length L, and the following can be written:

$$L = \int_0^\tau v \cdot \mathrm{d}t = \frac{eE}{2m^*} \cdot \tau^2. \tag{35}$$

If an average electron with a mean free path before the first collision of 0.1 μm is chosen, the value for τ is found to be

$$\tau = \sqrt{\frac{2m^*L}{eE}} \approx 8.7 \cdot 10^{-14}\,\text{s} \approx 0.1\,\text{ps}. \tag{36}$$

for a value of $E = 10^4$ V/cm. Therefore, the maximum velocity that the electron can acquire is given by

$$v_{max} = a \cdot \tau = \sqrt{\frac{2eEL}{m^*}} \approx 7.3 \cdot 10^7\,\text{cm/s} \tag{37}$$

Since the peak steady-state velocity of electrons in GaAs is in the range of $1.5 \cdot 10^7$ cm/s to $1.5 \cdot 10^7$ cm/s, the maximum velocity derived above is an "overshoot" velocity that an electron acquires when in a very short gate region with a large electric field.

Keeping in mind the cross-section of a MESFET, consider a device that has a cross-section as in Fig. 19. In this High Electron Mobility Transistor (HEMT), GaAs is not the only material that is used. There are a number of so-called hetero-junctions, i.e. semiconductor junctions between different materials. The most important one in terms of device operation is that between the silicon-doped AlGaAs and the undoped GaAs. Due to the higher band-gap of AlGaAs compared to GaAs, free electrons diffuse from the AlGaAs into the GaAs forming a two-dimensional electron gas at the interface. These electrons are confined to a very thin sheet because of the built-in potential barrier. It is easy to understand qualitatively why the transport properties of electrons in this region are superior to those in the channel of a MESFET: the MESFET channel must be doped to have current, and the electrons scatter off the dopant ions. In the thin layer of electons in the a HEMT, there are no ions to scatter off, so the electrons can gain very high velocities , i.e. their mobility is very high. This is somewhat of a subtle point: initially it was thought that the excellent properties of HEMTs are due to the high mobility of electrons (thus the initial name), but later it became clear that it is in fact the average high electron velocity that enables high frequency operation with superior noise figure.

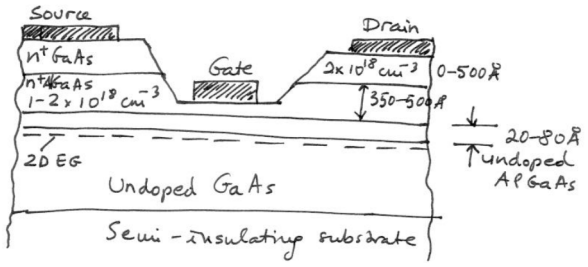
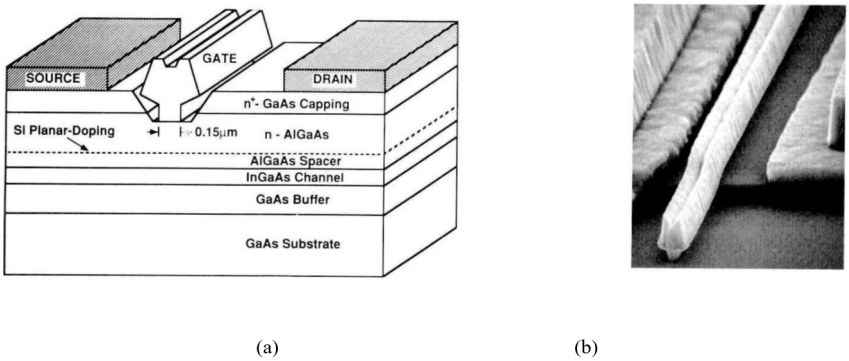


Fig. 19 Sketch of the cross-section of a HEMT device, with approximate dimensions of the different layers grown by MBE.

From Fig. 19 one can see that the layers of different semiconductors are extremely thin. Technology was not mature enough to enable such material structures until Molecular Beam Epitaxy (MBE) was invented in Bell Labs in the 1970's. The problems in growing such structures are associated with lattice mismatches between different semiconductor crystals. The first heterostructures were grown to investigate optical spectroscopy by a physicist from Bell Labs, R. Dingle. He wrote: "*...since multiple layers could be readily grown, we simply grew a multilayer AlGaAs/GaAs structure containing* 10 *or* 20 *layers interleaved with AlGaAs support layers. The growth technique was described as "semi-automatic" and consisted of watching the second had of a darkroom timer and manually rotating a shutter on the aliminum effusion oven of the MBE system it initiate and terminate AlGaAs layer growth. In early* 1974 *a multilayer structure with* 200*-angstrom thick GaAs layers and thicker AlGaAs support layers was gown. With the help of Len Kopf, we measured the absorption spectrum at* 2 *K and observed the first direct evidence for size quantization of electron motion in GaAs. There was great jubilation in my lab – we even danced a bit, as I recall*! *I began to believe in quantum mechanics*!" What was in effect observed was electron motion in the 2-D electron gas. A mobility of 10,000-20,000 cm^2/Vs was first measured at low temperatures, while the common bulk GaAs mobility was 6,000 cm^2/V. Eventually, as high as 2,000,000 cm^2/V was obtained at low temperature and 8,500-9,000 cm^2/V at room temperature.

When one looks through FET device specifications, one often runs into the acronym PHEMT. The "P" stands for "pseudomorfic" and what it means is that, in order to improve the performance of a HEMT, the two-dimensional electron gas is confined to a thin layer of InGaAs instead of GaAs. This allows for even higher sheet charge density of the 2-D electron gas, and therefore higher transconductance. The cross-section of a typical PHEMT, along with a SEM photo of the 0.15 μm T-gate structure is given in Fig. 20.



(a) (b)

Fig. 20 (a) Cross-section of a typical PHEMT, and (b) SEM photo of the 0.15 μm T-gate metal on top of a HEMT channel.

4 Introduction to Microwave Transistor Amplifiers

The general block diagram of a single-stage microwave amplifier is given in Fig. 21. The active device can be a FET or bipolar device, and this will determine the type of biasing. In a FET amplifier using a depletion-mode device, the gate is biased negatively w.r.t. the source, and the drain positively. The source is usually grounded, both RF and dc-wise. The negative gate bias usually needs to be turned on first to avoid burning the transistor, often referred to as "bias sequencing". In contrast, bipolar transistors do not require bias sequencing, so they can easily be biased with a single dc supply. This was one of the motivations for developing enhancement-mode FETs, for which the gate is biased positively w.r.t. the source, and at a lower bias voltage than the drain, to enable a single-polarity supply.

The bias is supplied through a biasing circuit that needs to present a high impedance to all present RF signals so as not to present an additional (usually not well characterized) load. This is relatively straightforward in a narrowband amplifier design, but becomes a challenge for the broadband case. RF capacitors are needed to block the DC signal to the RF input and output. Usually the source (or emitter) terminal of the active device is connected to RF (and often DC) ground. In the case of microstrip, one or more metalized via holes are used for grounding, and they present an equivalent inductance between the terminal and ground. You will examine the effect of this inductance in your project. In the case of coplanar waveguide (CPW) circuits, the connection is more straightforward and the parasitic reactance can be minimal.

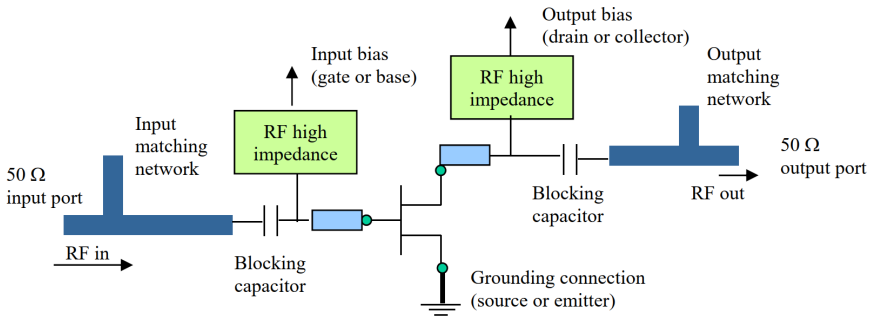


Fig. 21 General circuit diagram of a microwave amplifier.

Matching circuits at the input and output determine ultimately the performance of the amplifier. The following types of amplifiers result from different matching circuits:

- small signal gain-matched amplifier −− input and output are matched for best return loss;

- low-noise amplifier −− input is matched to a special impedance that cancels some of the noise originating from the amplifier input and output. Output port is matched for good return loss;
- broadband amplifier −− there are several architectures that enable broadband operation, usually they involve more than one active device. It is difficult to provide broadband matching to a single-stage amplifier without introducing substantial loss;
- high-power amplifier −− output is matched for large-signal maximum power delivery to the load, and input is matched to maximize gain;
- high-efficiency amplifier −− several options exist depending on other requirements.

In all the above designs, the first step is to ensure that the amplifier will be stable, i.e. that it will not be an oscillator. It is relatively straightforward to design stable small-signal amplifiers, since the stability criteria can be formulated in terms of transistor *S*-parameters. In saturated amplifiers (power amplifiers), small-signal transistor parameters are not valid, and stability is more difficult to predict during the design.

4.1 Small-Signal Linear Amplifier Design

Transistor *S*-parameters are usually given as common-source two-port parameters. Since there is feedback between the output and input port of a realistic transistor, the two-port is considered to be bilateral. In that case, the input scattering parameter is affected by the load impedance through this feedback (i.e. it is different from s_{11} of the device), and the output scattering parameter is affected by the generator impedance (i.e. it is different from s_{22} of the device). From Fig. **??**, the input coefficient of the two port, with given *S*-parameters, terminated in an arbitrary load is found from the reflection coefficient definition:

$$s_{in} = \frac{b_1}{a_1} = \frac{s_{11}a_1 + s_{12}s_L b_2}{a_1} = s_{11} + \frac{s_{21}s_{12}s_L}{1 - s_L s_{22}} \tag{38}$$

where b_2 is eliminated from the above expression using the relation:

$$b_2 = s_{11}a_1 + s_{22}a_2 = s_{11}a_1 + s_{22}s_L b_2 \tag{39}$$

Similarly, the output reflection coefficient looking into port 2 is found to be:

$$s_{out} = \frac{b_2}{a_2} = s_{22} + \frac{s_{21}s_{12}s_g}{1 - s_g s_{11}} \tag{40}$$

where s_g is the generator reflection coefficient. These expressions are used to define various transistor and amplifier gain values, as well as for stability analysis, as discussed in the book by Gonzalez and other references.

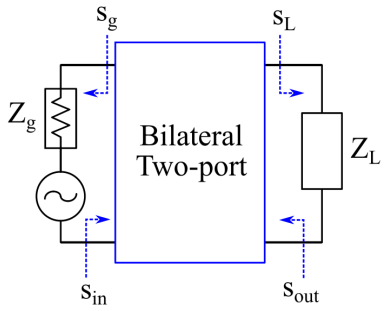


Fig. 22 Input and output scattering parameters of a bilateral two-port network.

Whether one is interested in buying or designing a small-signal amplifier (driver, gain stage), the following questions need to be answered:

1. What is the maximum and minimum required gain?
2. What is the operating frequency and bandwidth?
3. Is the amplifier matched and stable?

To answer these questions, we need to define a set of parameters that describe the amplifier. These parameters are defined for linear amplifiers with time-harmonic input signals. There are several definitions of the power gain for amplifier circuits. One definition useful in amplifier design is the transducer gain, $G_T = P_L/P_{av}$, where P_{av} is the power available from the generator. This definition of gain is a ratio of the power delivered to a matched load divided by the power that the source would deliver to a matched load in the absence of the amplifier. If the amplifier is not matched to the source, then reflected power is lost and a second definition, the power gain, $G_P = P_L/(P_{av} - P_{refl})$ can be used. Finally, if an amplifier is not matched at the load, the power available from the source will not be delivered to the load, so the available gain G_A is defined in terms of the power available from the amplifier network, $P_{av,N}$ as $G_A = P_{av,N}/P_{av}$. Expressions for these three gains in terms of the transistor S-parameters are found in most texts.

In this type of analysis, there are no assumptions about the device parameters. Sometimes, especially at lower frequencies, the internal feedback between output and input of the device is not significant, and the input and output matching can be done separately. This is referred to as *unilateral matching*. If the device can be considered as unilateral, this effectively means that s_{12} can be set to zero, but there will be some error involved in this assumption. A figure of merit, called the unilateral figure of merit, U, is defined to quantify this error as follows:

$$\frac{1}{(1+U)^2} < \frac{G_T}{G_{TU}} < \frac{1}{(1-U)^2}, \quad \text{where} \quad U = \frac{|s_{12}| \cdot |s_{21}| \cdot |s_{11}| \cdot |s_{22}|}{(1-|s_{11}|^2)(1-|s_{11}|^2)} \tag{41}$$

where G_{TU} is the unilateral transducer gain obtained when $s_{12} = 0$. The value of U varies with frequency, since the S-parameters vary with frequency. In order to check

if one can use the unilateral approximation at a given frequency, it is advisable to plot U and determine a range where its absolute value is smaller than approximately −15 dB (which corresponds to the ratio G_T/G_{TU} being smaller than about 0.25 dB).

Small-signal amplifier design starts from measured S-parameters at a particular drain voltage and current. Given that the bias and input drive level (small signal) have been chosen for us, our design task is to find input and output matching sections for the transistor. The assumption is that the manufacturers of the transistor were kind enough to give us the S-parameters of the device for the chosen bias point and input drive. For example, let us assume that we are given the following S-parameters at 5 GHz:

Frequency	s_{11}	s_{12}	s_{21}	s_{22}
5 GHz	$0.7\angle 100^\circ$	0	$2.5\angle 60^\circ$	$0.7\angle -30^\circ$

If we now assume 50-Ω source and load impedances, the task becomes to transform both s_{11} and s_{22} into 50-Ω. We add reference planes to the circuit diagram so that we can properly define the relevant circuit reflection coefficients. A common notation for the input and output reflection coefficients is s_{in} and s_{out}, where these represent the reflections off of the input and output of the transistor, respectively. In this unilateral example, they are equivalent to s_{11} and s_{22}. If s_g is the reflection coefficient looking into the input matching section toward the source and s_L is defined looking into the output match towards the load, to perform a conjugate match, we simply want design matching sections such that $s_g = s_{in}^*$ and $s_L = s_{out}^*$ This means that $s_g^* = 0.7\angle 100^\circ$ and $s_L^* = 0.7\angle -30^\circ$. By conjugate matching the input and output of the device to the source and load, we have guaranteed the maximum transducer gain possible for this amplifier. This matched transducer gain is a function of s_{21} and the matching circuits, and is given by:

$$G_{TU} = \frac{1}{1-|s_g|^2}|s_{21}|^2\frac{1-|s_L|^2}{|1-s_{22}s_L|^2} \tag{42}$$

We may stop here with the design of the unilateral case. We have achieved the best possible match and gain of the device at 5 GHz. Calculate the unilateral transducer gain, how does it compare to $|s_{21}|^2$?

We have neglected many things in this simple example:

- Usually s_{12} is not zero or negligibly small;
- The transistor in this circuit has behavior at other frequencies which needs to be analyzed;
- Maybe we are concerned with noise, power or efficiency in addition to gain.

If the error in G_T/G_{TU} is not smaller than about 0.25 dB, then a bilateral matching will be required. In this case, a conjugate match of the amplifier means matching to the conjugate of s_{in} and s_{out}, not s_{11} and s_{22}. Note that s_{in} is a function of s_L, the output loading, and s_{out} is a function of the source loading s_g. Therefore, a simultaneous match of the input and output is required (in other words, what we do at the input affects the output match, and vice versa).

If the s_{12} parameter is not negligible, the analysis becomes more complex, but we can still follow the strategy in the previous example. If $|s_{12}| > 0$, it is mostly due to the capacitance between the gate and drain discussed earlier in the equivalent circuit model of the device. The important result of this input-to-output coupling is that we can no longer say that $s_{in} = s_{11}$ as in the unilateral case. Instead, the input and output matching network need to perform conjugate matches as follows:

$$s_g^* = s_{in} = s_{11} + \frac{s_{21}s_{12}s_L}{1 - s_L s_{22}} \qquad \text{and} \qquad s_L^* = s_{out} = \frac{b_2}{a_2} = s_{22} + \frac{s_{21}s_{12}s_g}{1 - s_g s_{11}} \quad (43)$$

This result has several important implications:

1) A conjugate match of the amplifier means matching to the conjugate of s_{in} and s_{out};
2) s_{in} is a function of s_L, the output loading, and s_{out} is a function of the source loading. Therefore, we now have to perform a simultaneous match of the input and output (in other words, what we do at the input affects the output match, and vice versa).
3) The denominators $(1 - s_{22}s_L)$ and $(1 - s_{11}s_g)$ are < 1. This indicates that for some values of the transistor S-parameters, s_g, and s_L, it may be possible for $s_{in}, s_{out} > 1$, which indicates an instability.

Regarding 1) and 2), a simultaneous match can be found analytically by solving simultaneously the equations for the source and load reflection coefficients. The result is:

$$s_g^* = B_1 \pm \frac{\sqrt{B_1^2 - 4|C_1|^2}}{2C_1} \quad \text{and} \quad s_L^* = B_2 \pm \frac{\sqrt{B_2^2 - 4|C_2|^2}}{2C_2} \quad (44)$$

where $B_1 = 1 + |s_{11}|^2 - |s_{22}|^2 - |\Delta|^2$, $C_1 = s_{11}^2 - \Delta \cdot s_{22}^*$, $B_2 = 1 + |s_{22}|^2 - |s_{11}|^2 - |\Delta|^2$, $C_2 = s_{22}^2 - \Delta \cdot s_{11}^*$, and $\Delta = s_{11}s_{22} - s_{12}s_{21}$ is the determinant of the scattering matrix. *Note*: you can only use these formulas if the device is unconditionally stable! If it is not (in most cases), you can do a few things:

- Make it unconditionally stable by adding, e.g. a resistor in the gate, and recalculating the S-parameters of this "new" device;
- Use a simulator to slowly tune the input and output matches from the unilateral case while watching the trends in gain, stability, match, etc.

4.2 Stability

Note that the denominators $(1 - s_{22}s_L)$ and $(1 - s_{11}s_g)$ for some values of the transistor S-parameters, s_g, and s_L may result in in the expressions for $s_{in} > 1$ and $s_{out} > 1$, which indicates that the transistor amplifier circuit is giving power back

to the source. If a reflection coefficient becomes greater than 1, then oscillations are likely to occur, a phenomenon that we will exploit later in designing oscillators, but is an unwanted effect in amplifiers. Stability is often verified ahead of time, and there are several criteria that are commonly used, such as using stability circles of allowed and non-allowed impedances. The derivation of the stability circles is messy but straightforward (to see how it is done, look up, e.g. Gonzalez's book). Briefly, the condition for $|s_{in}| = 1$ and $|s_{out}| = 1$ are re-written as equations of circles, and points either inside or outside of those circles are values of load and source impedances that result in instabilities.

Alternatively, one can calculate two numbers at a given frequency point to determine if the transistor will be unconditionally stable, i.e. stable for all loads. The stability factor K, given by

$$K = \frac{1 - |s_{11}|^2 - |s_{22}|^2 + |\Delta|^2}{2|s_{12}s_{21}|} \tag{45}$$

along with the scattering matrix determinant $|\Delta|$ are calculated and compared to unity:

- If $K > 1$ and $|\Delta| < 1$, then the transistor is unconditionally stable at that frequency. This means that no matter what the load and source impedances are, s_{in} and s_{out} will always be less than unity.
- If $K < 1$ and $|\Delta| < 1$, the transistor is conditionally stable at that frequency. This means that for some values of the load and source impedances, s_{in} and s_{out} may exceed unity, in which case we need to do additional work when designing the matching sections, as will be discussed next week in class. When designing an amplifier, it is also necessary to check the stability at other frequencies. If there are instabilities, then it may be necessary to add resistors in the design. In this way, we may trade stability for some gain.

Which of the possible combinations of the solutions in the solutions to the simultaneous matched condition do we choose? It turns out that the two solutions with the negative sign in Eqs.(44) are the ones that are useful for an unconditionally stable network. If $B_1 > 1$ and $|B_1/2C_1| > 1$ in the equation for s_g, then the solution with the negative sign produces $|s_g| < 1$ and the solution with the plus sign results in $|s_g| < 1$ and similar statements are true for s_L. The maximal transducer power gain under simultaneously matched conditions can be written in the following format:

$$G_{T,max} = \frac{|s_{21}|}{|s_12|}\left(K - \sqrt{K^2 - 1}\right). \tag{46}$$

Under simultaneous matched condition, $G_{T,max} = G_{P,max} = G_{A,max}$. The maximal stable gain will be the value when $K = 1$, i.e. $G_{max,stable} = |s_{21}|/|s_{12}|$. This is a figure of merit for potentially unstable transistors and is often given in manufacturer's specification sheets.

Recommended reading related to this lecture: *Microwave Transistor Amplifiers*, G. Gonzales, 1984 edition, Chapter 1 and pages 92-132; and *Fundamentals of RF and Microwave Transistor Amplifiers*, Bahl, Chapters 1 and 2.

4.3 Stability with Source-Via Inductance – 2- to 3-port Conversion

We also talked about stability when source inductance is included in the model. When grounded-source two-port S-parameters are given, in order to disconnect the source from ground, add inductance, and convert to a new two-port network, the 2-port first needs to be converted to a 3-port. For a Y-matrix, under the assumption that KCL holds for the gate, drain and source currents, and referring to Fig. 23, we can obtain the relationship between the Y parameters of the 2-port network (Y) and the three-port network (Y'), as follows:

$$\begin{aligned} I_1 &= y_{11}(V_1 + V_3) + y_{12}(V_2 + V_3) \\ I_2 &= y_{21}(V_1 + V_3) + y_{22}(V_2 + V_3) \\ I_3 &= -I_1 - I_2 \end{aligned}$$

Rearranging, we can obtain:

$$Y' = \begin{bmatrix} y_{11} & y_{12} & -(y_{11}+y_{12}) \\ y_{21} & y_{22} & -(y_{21}+y_{22}) \\ -(y_{11}+y_{21}) & -(y_{12}+y_{22}) & -(y_{11}+y_{12}+y_{21}+y_{22}) \end{bmatrix} \tag{47}$$

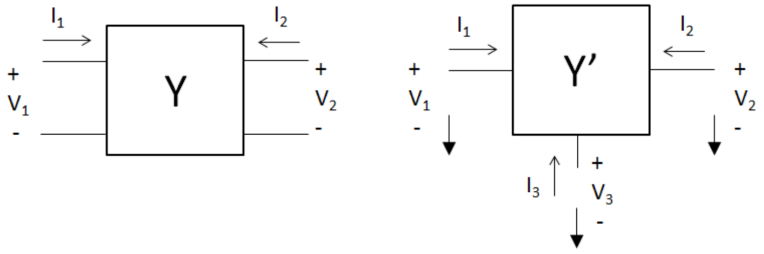


Fig. 23 Converting two-port Y-parameters to three-port Y-parameters.

5 Signal Flowgraphs for Microwave Circuit Analysis

In many texts on microwave amplifiers (e.g. the classic one by Gonzalez), signal flow-graphs are used for defining gain, determining stability, etc. This lecture gives a brief review of signal flow-graphs as they apply to microwave circuits, i.e. s-parameters and wave variables.

One of the main reasons for using signal flow-graphs is that, by following some simple rules, one can quickly find one of the parameters of the circuit, so that there is no need to waste time on solving for parameters that are not of interest. Signal flow graphs were developed a few decades ago for control theory, and were not used in circuit theory because circuit models are easier to solve. In microwaves, however, scattering parameters and incident and reflected waves can be represented easily with signal flow graphs. The classic paper on signal flow graphs was written by Samuel Mason at MIT ("Feedback Theory – Further Properties of Signal Flow Graphs," Proceedings of the Institute of Radio Engineers, volume 41, pp. 1144-1156, 1953).

Let us look at a linear equation for a one-port network, for example some load with a reflection coefficient s, Fig. 24(a): $b = sa$, where a is the incident wave, and b is the reflected wave. In signal flowgraphs, one thinks of waves as variables, and scattering parameters are constants. The variables become nodes of the flow graph and the constant coefficients become branches, Fig. 24(b). The arrow on the branch points from the variable on the right side of the equation to the variable on the left side. Let us now add a section of transmission line of electrical length in front of the one port load, Fig. 25(a). What does the signal flow graph for this network look like? Let us first write the scattering matrix for a transmission line section:

$$S = \begin{bmatrix} 0 & e^{-j\theta} \\ e^{-j\theta} & 0 \end{bmatrix} \tag{48}$$

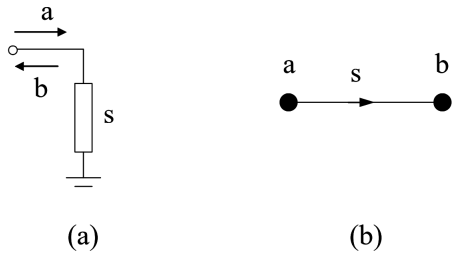


Fig. 24 A one-port network described with scattering parameters (a) and its signal flow graph (b)

Now the variables need to be labeled. The forward waves for the transmission line are the reflected waves for the load, so the following can be written, from Fig. 25(a):

$$b_2 = e^{-j\theta}\, a_1 \tag{49}$$

$$a_2 = s\, b_2 \tag{50}$$

$$b_1 = e^{-j\theta}\, a_2 \tag{51}$$

The signal flow graph is shown in Fig. 25(b). It is a cascade of three branches. The direction of the arrows is very important, since it shows which variables are on which side of the equations. Solving for b_1 gives:

$$\frac{b_1}{a_1} = e^{-j\theta}\, s\, e^{-j\theta}. \tag{52}$$

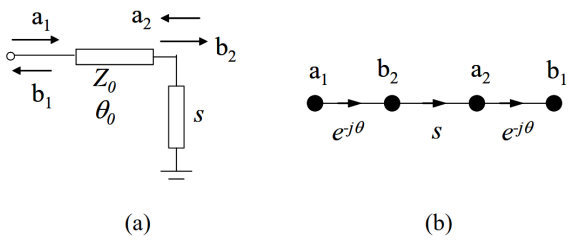


Fig. 25 A section of transmission line of impedance and electrical length connected to a load (a) and the signal flow graph (b).

From here we can see that the three branches form a path from a_1 to b_1. A path is a collection of branches and nodes that allow one to move from a beginning node to an ending node, following the arrows from node to node. The *gain* from a_1 to b_1 is given by the product of the coefficients of the three branches connecting these nodes. You can see from here that the gain of a circuit can be found easily by just multiplying cascaded branch coefficients with arrows in the same direction. When there is a loop in the graph, it modifies the gain.

Let us now look at a circuit that has a loop in the signal flow graph, Fig. 26(a). We will find the scattering parameter from port j of the network S to port i of the network T. The two connected ports are k on S and m on T. Let us call the combined network S' and find its parameters s'_{ij}. Combining smaller networks into larger ones is an important problem, as it is used in a number of circuit simulators, as will be illustrated at the end of this section.

From the definition of the scattering parameters, we can write expressions for the incident and scattered waves from Fig. 26(a) as:

$$b_k = s_{kj}\, a_j + s_{kk}\, a_k \tag{53}$$

$$b_m = t_{mm}\, a_m \tag{54}$$

$$b_i = t_{im}\, a_m \tag{55}$$

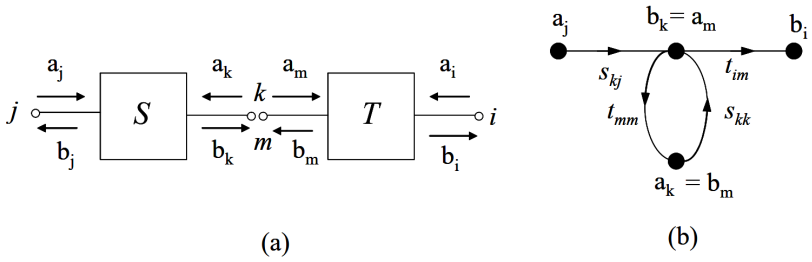


Fig. 26 Two two-port networks joint with one port (a) and the signal flow graph (b).

Further, since we know that port k is connected to port m, we have that $a_k = b_m$ and $b_k = a_m$. Now we can write a signal flow graph with four nodes and four branches, as shown in Fig. 26(b). The branches point from the variables on the right side of the equation to the variables on the left. Here we have a node with two branches entering and two branches leaving. The graph has a loop, which is a path that ends where it started. The gain of the loop is $s_{kk}t_{mm}$. A loop is said to touch another loop or path when it shares a node with it.

We can solve the equations written above to get:

$$s'_{ij} = \frac{s_{kj}\, t_{im}}{1 - s_{kk}\, t_{mm}} = \frac{b_i}{a_j}, \tag{56}$$

where you can see that the numerator is the gain of the path from a_j to b_i, and the denominator is 1 minus the loop gain.

When there are more than 1 path between two nodes, the contributions from all paths are added according to Mason's rule:

$$G = \sum_k G_k \frac{\Delta_k}{\Delta}, \tag{57}$$

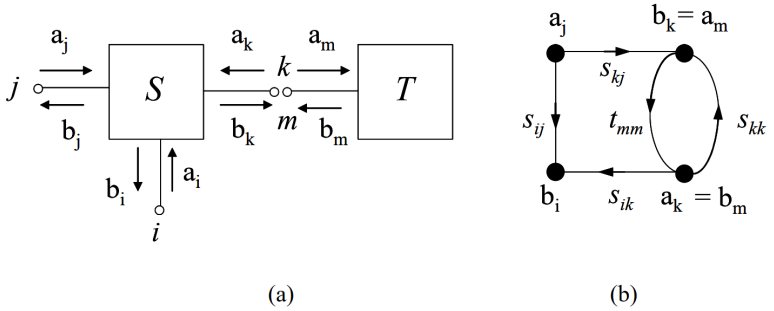
where G_k is the gain of path k, Δ is the determinant of the graph, and Δ_k is the cofactor of path k. The determinant of the graph is given by

$$\Delta = 1 - \sum_n P_{n1} + \sum_n P_{n2} - \sum_n P_{n3} + \ldots \tag{58}$$

where $\sum_n P_{nr}$ is the gain product of the n-th possible combination of r non-touching loops. The cofactor Δ_k of path k is the determinant of the loops that do not touch the path.

As an example, let us look at combining two networks, Fig. 27(a) and the associated signal flow graph, Fig. 27(b). Let us solve for s'_{ij}. There is a direct path between b_i and a_j, and also an indirect path with a loop that has a gain of $s_{kj}t_{mn}s_{ik}$, so from Mason's rule we get:

$$s'_{ij} = s_{ij} + \frac{s_{kj}\, t_{mm}\, s_{ik}}{1 - s_{kk}\, t_{mm}} \tag{59}$$



(a) (b)

Fig. 27 Combining two networks (a) and signal flow graph (b). The output and input ports are on the same network.

since $G_1 = s_{ij}$, $\Delta = 1 - s_{kk}\, t_{mm}$, $\Delta_1 = \Delta$, and $\Delta_2 = 1$

A number of circuit simulators use a method referred to as sub-network growth and signal flow graphs. Let us briefly discuss this on an example of the directional coupler, Fig. 28. First, the circuit is interpreted as eight parts: four transmission lines and four tees. Then, the parts are joined in pairs to make four new three-ports as shown in the figure. Next, two four-ports are formed, ABCD and EFGH, which are then joined to make a six-port. Finally, two of the six ports are joined. In the process, only two types of connections are solved, Fig. 29: when the ports are on different networks, and when the joining ports are on the same network. We have already solved case (a), where the joining ports are k and m. For practice, find the signal flow graph and solve for s'_{ij} for an internal connection, Fig. 29(b). This is a bit more complicated than the previous cases.

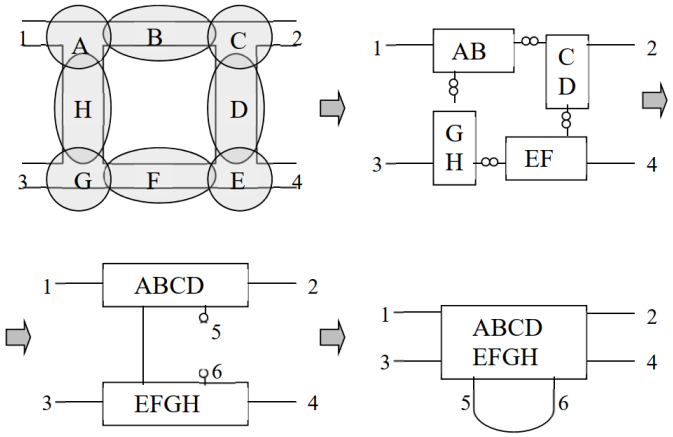


Fig. 28 Illustration of signal flow-graph applications to circuit simulations: subnetwork growth on the example of a branch line coupler.

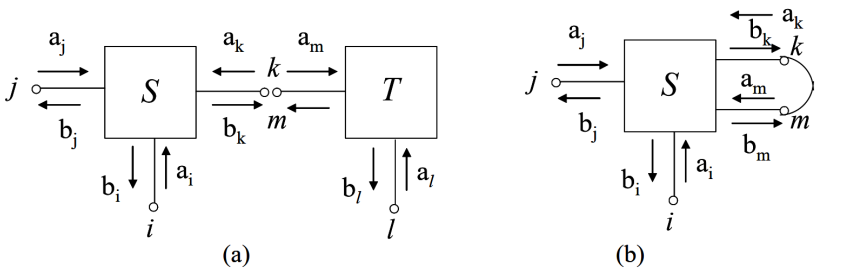


Fig. 29 A joint between ports k and m on two different networks (a) and on the same network (b).

5.1 Example: Coupled-Line Couplers

If you try to design a coupled-line coupler with a high coupling coefficient, such as 3 dB, you will find that the lines need to be too close and that it cannot be fabricated. For example, in the case of an alumina substrate, the gap between the two lines would need to be 10 μm, which is difficult to fabricate over a quarter-wave long section of line. To make a 3-dB coupled-line coupler that is shorter than a wavelength at the center frequency of the design, one can combine two or more couplers with lower coupling ratios as shown in Fig. 30(a) for the case of two couplers.

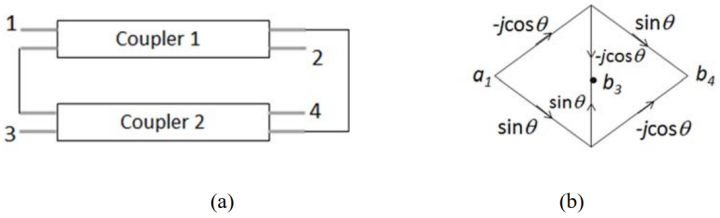


Fig. 30 (a) Connections for making a coupler with a larger coupling coefficient from two couplers with weaker coupling. (b) Signal flowgraph of (a).

Two coupled-line sections (each a 4-port) can be connected into a single coupler with a different coupling coefficient. A trick that makes it easy to solve for the coupling coefficient of couplers 1 and 2, assuming they are equal, is to write the coupled-wave S-parameters as: $\sin\theta$ for the coupled port transmission coefficient and $-j\cos\theta$ for the through-port transmission. They are in quadrature, as they should be for a symmetrical, reciprocal and lossless 4-port. We will next assume that the interconnecting lines are all of the same length to preserve symmetry. Now we can draw the signal flowgraph for the two connected couplers as shown in Fig. 30(b), and we write the S-parameters as:

$$s_{31} = \sin^2\theta - \cos^2\theta = -\cos(2\theta) \tag{60}$$

$$s_{41} = -2j\sin\theta\cos\theta = -j\sin(2\theta) \tag{61}$$

The coupled and through ports are still 90 degrees apart and port 2 is isolated. If we want the combined coupler to be a 3-dB coupler, we set:

$$\cos(2\theta) = \frac{1}{\sqrt{2}} \rightarrow \theta = 22.5^\circ \tag{62}$$

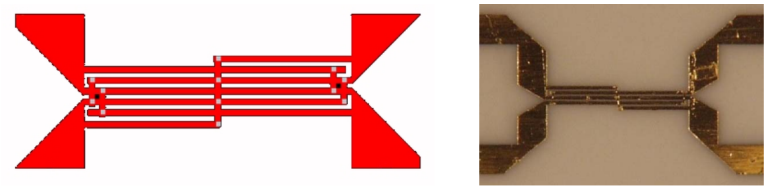
The coupling coefficient is now $\sin 22.5^\circ = 0.383$, or 8.343 dB. This means that by connecting two 8.343-dB couplers as shown in Fig. 30(a), we get a 3-dB 90-degree coupler with a large bandwidth.

The same idea is used in Lange couplers, which consists of very narrow multiple coupled lines of a quarter wavelength. A typical Lange coupler is shown in Fig. 31, and for the same substrate, the gap is 75 μm in comparison to the 10 mum that would be required for a single coupled section. The physical length of a Lange coupler is approximately equal to one quarter of a guided wavelength at the center frequency on the host substrate. The combined width of the strips is comparable to the width of a Z_0 (50-Ω) line on the host substrate.

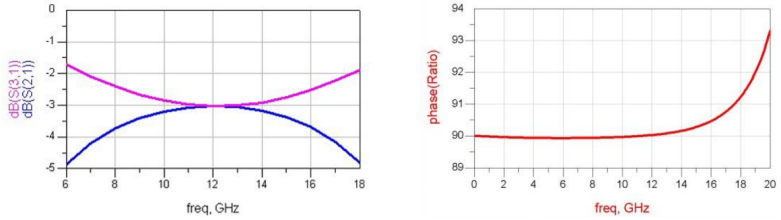
Lange couplers have been used from UHF to W-band, perhaps higher. However, as you go up in frequency, you will need to reduce your substrate height. Reduced height means reduced strip width, which is the ultimate limitation. At some point the strips get so narrow that even if they can be fabricated, they will become lossy. Lange couplers on alumina are usually restricted to applications where the substrate is 15 mils or thicker; this means you will see alumina Langes operate no higher than 25 GHz. If you attempted to make a Lange on 10-mil alumina, the strip widths would need to be less than 1 mil (25 microns). In MMIC applications, Lange couplers can be made on 4-mil and 2-mil substrates, although on 2-mil GaAs, the strip widths needs to be about five microns, but this is certainly within MMIC process precision.

Fig. **??**(b) shows the simulated response of an ideal Lange coupler in ADS. The match is below 20 dB for the entire range. The phase balance is flat with frequency, which is the main advantage of the Lange as a 90-degree coupler. Here is what Mr. Lange wrote about his invention:

"In 1969 *we at Texas Instruments were building microwave amplifiers on thin film ceramic substrates. We were using the scheme invented by Engelbrecht at Bell Labs, which required* 3*-dB quadrature couplers. The challenge was to get tight coupling on single layer microstrip. On the other hand our transistors had too much coupling between the interdigitated base and emitter fingers. So why not an interdigitated coupler*? *I built it*; *and it did not work well. Then I remembered that geometric symmetry guarantees quadrature, a* 90° *split between the outputs. So I moved some of the crossovers from the ends to the middle*; *and it worked*! *We had a microstrip interdigitated quadrature coupler with low loss and wide, one octave, bandwidth. "*



(a)



(b)

Fig. 31 Lange coupler layout (a). Referring to the six finger Lange, if the bottom left port is the input, the top left is the "coupled" port, the top right is the "through" port and the bottom right is the "isolated" port. You can find the "through" port easily in a Lange because it has a DC connection to the input. The isolated port is on the same side of the coupler as the input for a normal Lange. A photograph of an implemented Lange coupler on Alumina is shown on the right. (b) Simulated amplitude and phase balance using ADS for an ideal Lange coupler.

6 Broadband Amplifiers

The challenge in designing a broadband microwave amplifier is the fact that the input impedance at lower frequencies is practically an open circuit, and at higher frequencies predominantly capacitive and can be almost a short. This makes broadband matching difficult. (Check: how different is the impedance of the GaN HEMT you used in Project 2 at 200 MHz and 10 GHz?)

There are several ways to design a broadband input match for an amplifier, each has its drawbacks and advantages: (1) broadband non-uniform impedance matching network design; (2) balanced amplifier; (3) resistive feedback amplifier; (4) distributed (and a version called the traveling-wave amplifier, not to be confused with a TWT tube). The first type is based on designing dispersion-compensation or predispersion networks and results in very large impedance matching networks which typically have substantial loss.

6.1 Balanced Amplifiers

A common approach to the problem of broadband amplifier design is a balanced amplifier configuration shown in Fig. 32. It consists of a pair of 3-dB couplers.

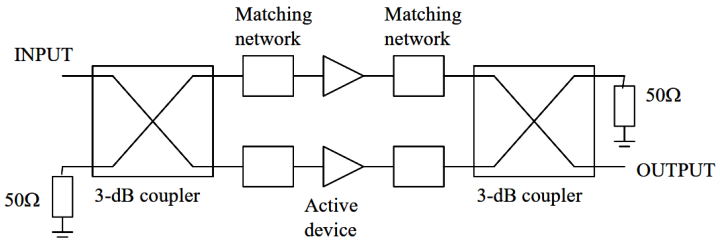


Fig. 32 Balanced amplifier configuration using 90-degree directional couplers.

For example, if the couplers are ideal hybrids (such as a branch line coupler), the scattering matrix for the balanced amplifier can be written as:

$$S = \frac{1}{2}\begin{bmatrix} (s_{11}^1 - s_{11}^2) & j(s_{12}^1 - s_{12}^2) \\ j(s_{21}^1 - s_{21}^2) & (s_{22}^1 - s_{22}^2) \end{bmatrix} \tag{63}$$

and if the two amplifiers are identical, the scattering matrix becomes:

$$S = \begin{bmatrix} 0 & js_{12} \\ js_{21} & 0 \end{bmatrix} \tag{64}$$

This means that, as long as the amplifier circuits are identical, they can be whatever we wish, and the amplifier still has input and output matching. The two amplifiers can individually be tuned for gain, noise or flatness of frequency response. Another common balanced amplifier uses Wilkinson combiner/dividers instead of the branch-line couplers. In this case, added quarter-wave sections in the two lines provide a 180-degree phase difference between the two waves reflected from the inputs of the amplifiers, and the reflections are cancelled. The bandwidth of the amplifier is obviously limited by the bandwidth of the directional coupler or Wilkinson splitter, both of which rely on quarter-wave sections for proper operation. Therefore, a hybrid is not the best choice (it has about 15% bandwidth). Instead, most commonly used is a Lange coupler based on coupled line sections, which can have a bandwidth of 2 octaves. It is also possible to design broadband multi-section branch line and Wilkinson combiners (over decade bandwidth).

Balanced amplifiers ideally have the same gain and twice the output power as compared to the single amplifier. When the signal becomes large, each of the transistors receives only half the power, so balanced amplifiers can handle more power with less signal distortion. However, twice the input signal is required, and two times more DC power. An additional disadvantage is the size of the circuit and the fact that a large part of the real-estate is taken by passive circuits. This is very costly in MMIC implementations.

An important factor in balanced amplifier design is the amplitude and phase mismatch between the coupler output ports as a function of frequency, as well as the sensitivity of this mismatch to load impedance variations. In a practical design, this should be verified in simulation prior to fabrication.

6.2 Distributed Amplifiers

A technique which achieves extremely broadband operation is the distributed amplifier, shown in Fig. 33. The idea behind it is that, instead of trying to tune out the transistor capacitances, these capacitances are used as part of a lumped-element approximation to a transmission line. The idea goes back to Percival in the 1930s (with a British patent in 1936), implemented in tube technology. Monolithic distributed amplifiers were demonstrated first in the early 1980's.

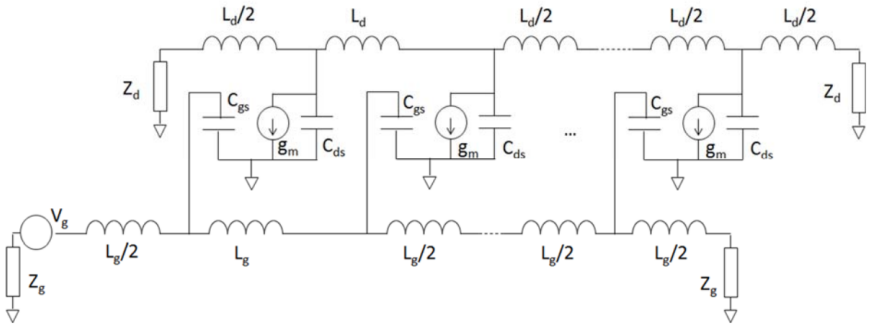


Fig. 33 A distributed MESFET amplifier using a very simplified unilateral FET model.

Consider the simplified equivalent circuit in Fig. 33. On the input side, inductors L_g are placed between the gate-to-source capacitances C_{gs} of the adjacent transistors, and in that way the familiar lumped-element artificial transmission line with a characteristic impedance of $Z_g = \sqrt{L_g/C_{gs}}$ is formed, and this impedance is nearly frequency independent. The phase velocity of a wave traveling along this line is $v_g = 1/\sqrt{L_g C_{gs}}$. This transmission line can be resistively terminated at the end with little loss of input signal. On the output side, inductors L_d are placed between drain-to-source capacitances of the adjacent devices, and a transmission line with a characteristic impedance $Z_d = \sqrt{L_g/C_{gs}}$ and phase velocity $v_d = 1/\sqrt{L_g C_{gs}}$ is formed. This is an active transmission line and the signal builds up along it. The phases of the outputs of the individual transistors will only be appropriate for left-to-right propagation, so little power will be lost in the resistive termination at the left end of the line. In effect, the two transmission lines are coupled lines with a coupling coefficient greater than unity. For a given transistor, the inductors L_g and L_d can be chosen to equalize the phase velocities of the two coupled lines. This discussion is of course valid only for a unilateral transistor approximation.

From this description of the distributed amplifiers, it appears that an arbitrarily large gain can be achieved by making a large number of sections. In the equivalent circuit for the transistor, however, there are some resistors as well, and this will make the transmission line lossy. As a result, a limited number of transistors can be added before the loss overcomes the gain. The frequency curve of the gain as a function of

the number of sections is shown in Fig. 34. It shows that after 5 sections, there is no appreciable increase in gain, whereas the flatness of the gain is reduced. Distributed amplifiers have been reported with flat gain from 1 to 40 GHz, and into the 100-GHz range.

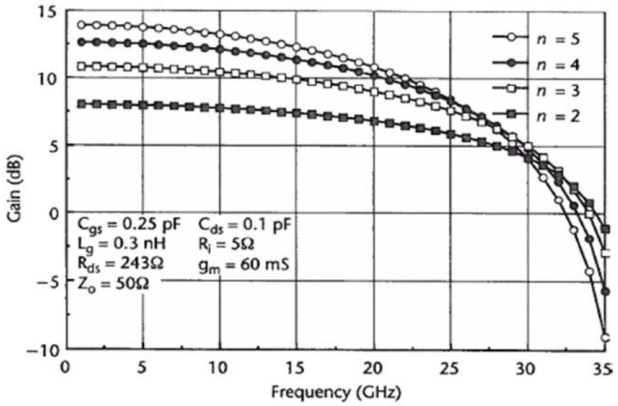


Fig. 34 Dependence of the gain versus frequency as a function of the number of sections of a distributed amplifier (from published data).

Distributed amplifiers are monolithically integrated so that the devices are very small compared to the guided wavelength. In practice, it is difficult to make good inductors in monolithic circuits (why?) at high microwave frequencies. Therefore, short sections of transmission lines are used instead between the stages of a distributed amplifier. Since in that case, the artificial transmission line model becomes even more of an approximation, these amplifiers are often viewed as traveling wave devices. Many commercially-available distributed amplifiers have a cascode configuration with two transistors in each cell. This helps boost the gain and extends the gain-bandwidth product.

Some advantages of distributed amplifiers include:

- Good input match, therefore easy to cascade;
- High isolation between output and input (typically better than 40dB), so stability is good;
- Current combines at the output, so power increased;
- Relatively insensitive to variations in device characteristics; and
- The noise figure can be reduced when the number of devices is increased.

Disadvantages include the challenges associated with matching the drain and gate line phase velocities, which might require adding additional capacitances. Additionally, losses in the gate line limit the achievable gain and number of cells. Failure of any stage has a major effect on amplifier performance although the degradation is graceful.

6.3 Resistive Feedback Amplifiers

Resistive feedback can also be used for designing a broadband amplifier. The effect of a feedback resistor between the gate and drain of a FET is to lower the input and output impedance and to broaden the gain curve. The drawback is resistive coupling between the bias circuits, as well as overall lower gain than for reactively matched amplifiers.

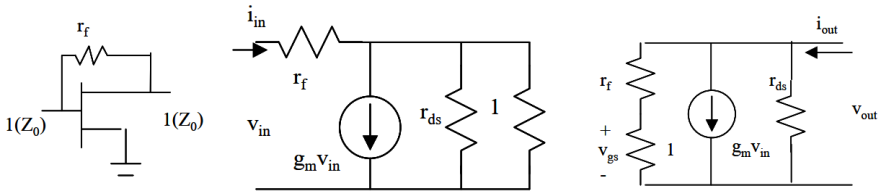


Fig. 35 Simplified equivalent circuits for input and output MESFET circuits in a series feedback amplifier.

By observing the approximate circuit model for the MESFET/HEMT with normalized resistance values (to 50 Ω), Fig. 35, the equations for the current and voltage in the input circuit are found to be:

$$i_{in} = g_m\, v_{in} + i \tag{65}$$

$$v_{in} = r_f\, i_{in} + \frac{r_{ds}}{r_{ds}+1}\, i \tag{66}$$

Here all the resistors are normalized to Z_0, so $r_{ds} = R_{ds}/Z_0$, etc. Solving for the input impedance by eliminating i, we obtain:

$$Z_{in} = \frac{v_{in}}{i_{in}} = \frac{r_f + \frac{r_{ds}}{r_{ds}+1}}{1 + g_m \frac{r_{ds}}{r_{ds}+1}} \tag{67}$$

For the output circuit, the current and voltage can be expressed as:

$$i_{out} = \frac{v_{out}}{1+r_f} + g_m\, v_{gs} + \frac{v_{out}}{r_{ds}} \tag{68}$$

$$v_{out} = (1+r_f)\, v_{gs} \tag{69}$$

where the voltage v can be eliminated to give the expression for the output impedance:

$$Z_{out} = \frac{v_{out}}{i_{out}} = \frac{r_{ds}\,\frac{1+r_f}{1+g_m}}{r_{ds} + \frac{1+r_f}{1+g_m}}$$

In both cases, the impedance can be controlled by the amount of feedback resistance.

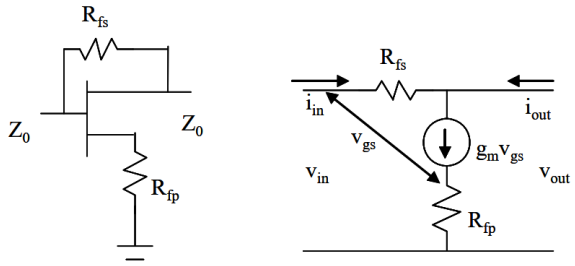


Fig. 36 Simplified circuit for series-shunt resistive feedback amplifier.

The above simplified analysis was an example of the more general case of series-shunt resistive feedback shown in Fig. 36, where in addition to the series feedback between gate and drain, there is a parallel feedback resistor placed in the source. The admittance matrix for this network (for a very simplified FET model) can be written as

$$\begin{bmatrix} i_{in} \\ i_{out} \end{bmatrix} = \begin{bmatrix} \frac{1}{R_{fs}} & -\frac{1}{R_{fs}} \\ \frac{g_m}{1+g_m R_{fp}} - \frac{1}{R_{fs}} & \frac{1}{R_{fs}} \end{bmatrix} \cdot \begin{bmatrix} v_{in} \\ v_{out} \end{bmatrix}. \tag{70}$$

The admittance matrix can now be converted to S-parameters using the standard conversion formulas:

$$S = \frac{1}{\Delta} \begin{bmatrix} 1 - \frac{g_m Z_0^2}{R_{fs}(1+g_m R_{fp})} & \frac{2Z_0}{R_{fs}} \\ \frac{-2g_m Z_0}{1+g_m R_{fp}} + \frac{2Z_0}{R_{fs}} & 1 - \frac{g_m Z_0^2}{R_{fs}(1+g_m R_{fp})} \end{bmatrix} \tag{71}$$

where

$$\Delta = 1 + \frac{2\,Z_0}{R_{fs}} + \frac{g_m Z_0^2}{R_{fs}(1 + g_m R_{fp})}$$

$$s_{11} = s_{22} = 0$$

If the design attempts to obtain a match at input and output, i.e. $s_{11} = s_{22} = 0$, then the resistor values are related to the transconductance by

$$1 + g_m R_{fp} = \frac{g_m Z_0^2}{R_{fs}} \quad \text{or} \quad R_{fp} = \frac{Z_0^2}{R_{fs}} - \frac{1}{g_m} \tag{72}$$

From the above equations, now S_{21} and s_{12} can be found to be

$$s_{21} = \frac{Z_0 - R_{fs}}{Z_0} \quad \text{and} \quad s_{12} = \frac{Z_0}{R_{fs} + Z_0} \tag{73}$$

Notice that the gain of the amplifier depends only on the characteristic impedance and the value of the series feedback resistor, not on the device parameters. This means that flat gain over a frequency range can be obtained with feedback.

The physical meaning of the above equations is that the input VSWR can be unity with a positive value of the parallel feedback resistor, if the transconductance of the active device is large. This is usually not the case in a MESFET/HEMT, but is the case in a bipolar transistor. For example, if we desire that the amplifier have , the minimal transconductance in a 50-ohm system and the value of the series feedback resistor are found by setting (this is the case that was first discussed with just the series feedback):

$$g_{m,\min} = \frac{1 - s_{21}}{Z_0} = 83\,\text{mS} \quad \text{and} \quad R_{fs} = 208\,\Omega. \tag{74}$$

This is a fairly large value of transconductance. Off course, the standard feedback relation $R_{fs} = Z_0(1 + |s_{21})$ is valid. When both series and shunt feedback resistors are used, and the transconductance is large enough, the best input and output match are obtained for $R_{fs}R_{fp} \approx Z_0^2$. This ignores the phase of s_{21}, which can vary rapidly as the frequency increases and cause positive feedback through the resistor. This can be solved by adding an inductor in the series feedback branch with a value that the amount of feedback decreases after a certain frequency.

7 Low-Noise Amplifiers

7.1 Introduction

Electrical noise is a random voltage or current which is present in a circuit with or without the presence of a signal. Usually, noise is unwanted. Noise should not be confused with interference, which is a signal coupling from another circuit, or with fading, which is random variations in the propagation characteristics in a radio link. Usually, the instantaneous values of the noise currents and voltages cannot be predicted, so their average values are zero. The average power, however, is not zero, and this is what noise is characterized by.

Noise in microwave devices can be caused by different physical phenomena: *thermal* (also referred to Johnson or white) noise associated with a resistor at some temperature; *flicker* or 1/f noise associated with low frequency variations, *shot noise* due to fluctuations in particle current in any device with a dc current flow, and *diffusion* noise produced by carriers in semiconductors which move due to diffusion.

Flicker noise has a power per unit bandwidth which varies with frequency as $1/f^{\alpha}$, where α is close to unity. Typically, it is important in the frequency range from much less than 1 kHz to MHz, depending on the device. For example, bipolar transistors in general have lower $1/f$ noise than FETs, so for low-phase noise oscillators, one should use bipolar devices. As the frequency increases beyond a corner frequency,

thermal noise starts dominating. There also must be a frequency in the low frequency limit where the noise becomes frequency independent, otherwise the integrated noise power would be infinite. This type of noise is important in oscillators, so we will revisit it later.

Shot noise is due to dc current in devices. Any current consists of discrete charged particles, with a number of particles per second given by $N = I_{dc}/e$ (e -– charge of electron), and statistical standard deviation of $\sqrt{N}$. Therefore, the noise current magnitude is

$$|i_n| \cong e\sqrt{N} \cong e\sqrt{\frac{I_{dc}}{e}}, \tag{75}$$

and since N is given per second, the bandwidth is 1 Hz. The power from the noise current in a bandwidth B is:

$$< i^2 >= 2e\, I_{dc}\, B, \tag{76}$$

where the factor of 2 appears because noise is always measured in a finite bandwidth. For a more detailed treatment of shot noise, a good reference is *Microwave Semiconductor Devices* by S. Yngvesson, Kluwer Academic Publishers, 1991.

The most significant noise at microwave frequencies is thermal noise. Thermal noise in a circuit is closely related to black body radiation, and can be understood starting from thermodynamic principles. Below is a discussion on thermal noise, which follows the classic paper "Thermal and quantum noise" by Oliver, IEEE Proceedings, May 1965.

7.2 Thermal Noise Spectral Power Density

In the derivation of thermal noise power per unit bandwidth (noise spectral power density), we start from the first and second law of thermodynamics and Plank's law: (1) in any closed system, the total energy is constant; (2) the entropy is maximized; and (3) electromagnetic energy is radiated and absorbed in discrete quanta (photons) of energy hf, where $h \approx 6.626 \cdot 10^{-34}$ Js is the Plank constant. Consider a lossless transmission line of length l, terminated in matched resistive loads, Fig. 37. If the two resistors are at different temperatures, the hotter one will lose energy to the colder one. We assume that the power is only on the transmission line (no radiation, convection or conduction). The question we want to answer is what noise power does the resistor give?

Imagine the resistors are shorted, trapping the thermal energy on the line. The line becomes a resonator that can support many modes. These modes are thermally excited in our scenario. To find the spectral power density (in W/Hz) of the thermal noise, we need to answer the following questions:

- How are the modes excited?
- How many modes are excited?
- How much energy does each mode contain?

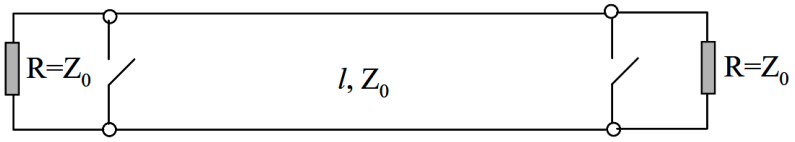


Fig. 37 A lossless transmission line l long, terminated in matched resistive loads $R = Z_0$. The resistors give rise to thermal noise power, which travels on the transmission line from one load to the other. The transmission line and loads are assumed to be a thermally isolated system.

To answer how resonant modes can be thermally excited, we start from the Boltzman H-theorem, which says the following. If a system can assume any of N discrete modes, each having a probability $p(n)$, as the closed system approaches equilibrium, the quantity

$$H = -\sum_{n=1}^{N} p(n)\ \log p(n)$$

never decreases, but tends to increase until the system reaches equilibrium. Here, the logarithm is in base e (so ln, but there are too many other n's so we will write it as log). The mathematical quantity H is directly related to physical entropy as $S = kH$, where $k = 1.3 \cdot 10^{-23}$ J/K is the Boltzman constant.

Consider now an electrical resonator at thermal equilibrium. The resonator can gain or lose energy only in discrete amounts, so at any moment there will be nhf energy in the resonator. Here n is such that the probability distribution $p(n)$ maximizes entropy. What is this probability distribution function equal to? To answer this question, we will (1) first find $p(n)$ for a given S_{max} and $\bar{n}$ (average number of modes), and then (2) determine $\bar{n}$ (adjust the mean) to maximize system entropy.

The entropy of the resonator is given by

$$S_r = k\, H_r = -k \sum_{n=0}^{\infty} p(n)\ \log p(n), \tag{77}$$

since there are an infinite number of modes in a resonator. For S_{max}, we set $\partial S_r / \partial p(n) = 0$, with the following constraints:

$$\sum_{n=0}^{\infty} p(n) = 1, \quad \text{where } n \text{ is an integer}$$

$$\sum_{n=0}^{\infty} n\, p(n) = \bar{n}, \quad \text{where } \bar{n} \text{ does not have to be an integer.}$$

What happens for small perturbations in $p(n)$? The two sums above do not vary, for a given average number of modes $\bar{n}$, and when $S_r = S_max$ (around the extremum), the sum $\Sigma p(n) \log p(n)$ will also not vary with small variations of $p(n)$.This in turn means that any linear sum of these three expressions will not vary for small deviations

$\partial p(n)$. Now we "fabricate" a quantity U that we know will have a zero variation with $\partial p(n)$:

$$U = \sum_{n=0}^{\infty} p(n)\, \log p(n) + \alpha \sum_{n=0}^{\infty} n\, p(n) + \beta \sum_{n=0}^{\infty} p(n). \tag{78}$$

We set $\delta U/\delta p(n) = 0$, and the following results:

$$\delta U = \sum_{n=0}^{\infty} [(\log p(n) + 1) + \alpha n + \beta] \cdot \delta p(n) \quad \Rightarrow \quad \log p(n) + 1 + \alpha n + \beta = 0$$

$$p(n) = e^{-1-\beta}\, e^{-\alpha n} = K\, u^n, \quad \text{where } K = e^{-1-\beta} \text{ and } u = e^{-\alpha}.$$

Now, K and $\bar{n}$ can be found from the constraints:

$$\sum_{n=0}^{\infty} K\, u^n = 1 \Rightarrow K = u - 1$$

$$\sum_{n=0}^{\infty} n\, p(n) = \sum_{n=0}^{\infty} n\, (1-u)\, u^n = (1-u)\frac{u}{(1-u)^2} = \bar{n}$$

This means that, since $\bar{n} = u/(1-u)$, adjusting u adjusts the mean $\bar{n}$. The probability distribution function for the number of modes (that should maximize entropy) is

$$p(n) = (1-u)\, u^n$$

The next question we need to answer is: What is u equal to? Let us first look at the evolved expression for the entropy, inserting the $p(n)$ form obtained above:

$$\begin{aligned} S_r &= -k \sum_{n=0}^{\infty} (1-u)\, u^n \log[(1-u)\, u^n] = \\ &= -k \sum_{n=0}^{\infty} [(1-u)\, u^n\, \log(1-u) + (1-u) u^n n \log u] = \\ &= -k\, (1-u) \log(1-u) \sum_{n=0}^{\infty} u^n - k(1-u) \log u \sum_{n=0}^{\infty} n u^n = \\ &= -k \left[\log(1-u) + \frac{u}{1-u} \log u \right] \end{aligned}$$

Imagine the resonator has no energy ($n = 0$) at $t = 0$ and that then the entropy is produced by draining an energy from the rest of the system. If the system is at a temperature T, this produces an entropy change (end minus beginning state) equal to

$$\Delta S = -\frac{\bar{n}hf}{T} = -\frac{u}{1-u}\frac{hf}{T}$$

in the rest of the system. In thermal equilibrium,

$$S = S_r + \left(-\frac{u}{1-u}\frac{hf}{kT}\right) = -k\left[\log(1-u) + \frac{u}{1-u}\left(\log u + \frac{hf}{kT}\right)\right]$$

the entropy is maximal, and $\partial S/\partial = 0$. Setting the derivative with respect to u of the previous equation, we obtain the following:

$$\frac{hf}{kT} + \log u = 0, \qquad u = e^{-hf/kT}, \tag{79}$$

$$p(n) = \left(1 - e^{-hf/kT}\right) e^{-hf/kT}. \tag{80}$$

The actual energy in the resonator is $nhf = W$. The probability of a state (mode) falls off as $e^{-W/kT}$ if we do not consider quantization. Quantization of modes only makes certain discrete modes possible. As frequency increases, the available levels become fewer and have more energy. When $hf/kT << 1$, the energy levels become less numerous and closely spaced, and then a continuous $p(n) = q(W)$ can be assumed. In the limit $hf \to 0$, the Boltzman distribution is obtained:

$$q(W) = \frac{1}{kT}e^{-W/kT}. \tag{81}$$

The average energy in the resonator is

$$\overline{W} = \bar{n}hf = \frac{u}{u-1}hf = \frac{hf}{e^{hf/kT}-1}. \tag{82}$$

Now we need to relate the energy trapped on the shorted transmission line to thermal noise power produced by the resistors, Fig. 37. There will be some mode density, i.e. number of modes per unit length of line per frequency, denoted by m_1, where the index "1" stands for one dimension since the transmission line is considered to be one-dimensional. Each of these m_1 modes in thermal equilibrium has an average energy $\bar{W}$, so the thermal energy density ρ_1 is the product of the mode density and the average mode energy: $\rho_1 = m_1\bar{W}$ (in J/m per unit bandwidth). As the line is made longer, the spectral density of modes increases in proportion to the length, so the energy density along the line is not changed. Half of the energy propagates in the $+z$, and half in the $-z$ direction.

The spectral mode density m_1 is found as follows. The modes on a shorted line ℓ long are given by $\ell = n\lambda/2 = (nc)/(2f)$, where c is the propagation velocity on the line. When the line gets very long, the number of modes does not depend on the frequency. How many modes per unit length per unit bandwidth are there? The answer is

$$m_1 = \lim_{\ell\to\infty}\frac{1}{\ell}\frac{dn}{df} = \frac{2}{c}. \tag{83}$$

Next we need to remember that the transmission line is in thermal equilibrium. If the shorts are now replaced by a matched load $R = Z_0$, the power density must remain unaffected. Otherwise, the load would either deliver or absorb energy, but it cannot do so in thermal equilibrium. Therefore, a matched load both absorbs and radiates a power $\bar{W}$. The two loads are exchanging energy at this rate and now the system is not resonant (the line length does not play a role). If one of the loads is cooled to 0 K, it will not produce noise power, and the energy flow from the hot load for $hf/kT << 1$ will be

$$W = \lim_{\frac{hf}{kt}\to 0} \frac{hf}{e^{hf/kt} - 1} \cong \frac{hf}{1 + \frac{hf}{kT} - 1} = kT \tag{84}$$

To obtain the power spectral density in the limit $hf/kT << 1$, the following reasoning applies:

- The energy spectral density W is the equal to the mode spectral density times the average mode energy: $W = m_1 \bar{W} = 2kT/c$.
- To obtain the power on the line, this energy density (in J/m/Hz) is multiplied by the velocity: $p = Wc = 2kT$ (in W/Hz).
- The above needs to be divided by 2 for the power flowing in only one direction down the line, so finally:

$$p = kT \tag{85}$$

is the power spectral density (W/Hz) of a resistor at temperature T for $hf/kT << 1$. The important practical implications of this result are:

- The power spectral density from a resistor due to random motion of electrons is proportional to the resistor temperature.
- $p \neq p(f)$ for the approximation in Eq.(85), which is why thermal noise is often referred to as white noise (at lower frequencies).
- At room temperature, $T = 290$ K, $kT = 4 \cdot 10^{-21}$ W/Hz. The noise power is given by kTB, where B is the bandwidth of the measurement/system. As the bandwidth of a system increases, this power does not go up to infinity, because Eq.(85) is only valid for a certain frequency range. Instead, the total thermal noise power available from a resistor from dc to infinite frequency is equal to, in watts,

$$P_R = \int_0^\infty \frac{hf}{e^{\frac{hf}{kT}} - 1} df = \frac{(kT)^2}{h} \int_0^\infty \frac{\alpha}{e^\alpha - 1} d\alpha = \frac{(kT)^2}{h} \frac{\pi^2}{6} \mathbf{W}.$$

- At $T = 290$ K, this power is equal to $P_R \approx 4 \cdot 10{-}8$ W, which is on the order of 10 nW. This is the rate at which the resistor would cool through electromagnetic radiation over a matched transmission line to a matched load at $T = 0$ K.
- If both terminations in Fig. 37 are at temperature T, the total noise power is $2P_R$, and the total RMS noise voltage on the line is $V_{noise} = \sqrt{2P_R R} = 2$ mV at room temperature for a 50 Ω resistor.
- For an unmatched load, with input impedance Z_{in}, in a certain bandwidth Δf, the noise power delivered to the unmatched load and the noise voltage are given by

$$P = \frac{|V_{\text{thermal noise}}|^2}{|R + Z_{in}|^2} R_{in} = \frac{2\,k\,T\,R\,\Delta f}{|R + Z_{in}|^2|} R_{in} \qquad \text{and} \qquad V_{noise} = \sqrt{2\,k\,T\,\Delta f\,R_{in}}.$$

- For active networks, the above reasoning is not valid, since the power is supplied from an external source, violating thermal equilibrium.
- The thermal noise from a resistor is the one-dimensional version of black body radiation, and the result is that there is no frequency dependence for many practical frequency ranges. In three dimensions, the mode spectral distribution can be shown to be a function of frequency, so the spectral power density will also be a function of frequency.

An interesting calculation is to find the frequency when the approximation in Eq.(85) stops being valid, e.g. there is an error of 3 dB. *Do this calculation as an excercise and also to know practically when Eq.*(85) *should not be used. Repeat the calculation for room temperature and liquid nitrogen* (77 K) *and liquid helium* (4 K).

7.3 Low-Noise Amplifiers

Low-noise amplifiers (LNAs) are a part of every radio receiver and typically follow the antenna. In some systems, there is a switch or a circulator (or both) between the antenna and the LNA and it is important to know how these affect the overall sensitivity of the receiver. We will first define the circuit parameters useful for LNA design, and then we will talk about how the noise that a transistor produces is described and related to these circuit parameters. Finally, we will discuss briefly how transistor noise parameters are measured.

7.3.1 Circuit Noise Parameters: Noise Figure and Noise Temperature

The noise figure of an amplifier or mixer give information about the noise that the amplifier adds to a signal while amplifying it, or the noise that the mixer adds to the signal while performing the frequency conversion. If the input to an amplifier is a signal S_i with a noise level N_i present, and at the output the signal and noise are S_o and N_o, the noise factor of the amplifier, Fig. 38 is defined as:

$$F = \frac{S_i/N_i}{S_o/N_o} \tag{86}$$

The Noise Figure (NF) is the noise factor expressed in dB: NF=10logF.

This quantity describes the degradation of the signal-to-noise ratio from input to output and is usually defined with the input noise being white noise with spectral power density of kT_0 at $T_=$290 K. A perfect amplifier would have a noise factor of 1 (NF=0 dB). A noise factor of 2 (NF=3 dB) means that the SNR has been reduced by a factor of two. A typical low-noise amplifier (LNA) you will design will have a

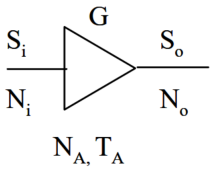


Fig. 38 Quantities used to define noise factor (figure) and noise temperature of an amplifier.

noise figure around 1–2 dB. By including the gain of the amplifier from Fig. 38, the following can be written:

$$S_o = G\,S_i \quad \text{and} \quad N_o = G\,N_i + N_A^o,$$

where N_{Ao} is the noise added by the amplifier measured at the output of the amplifier. The noise figure can now be written as

$$F = 1 + \frac{N_A^o/G}{N_i} = 1 + \frac{N_A}{N_i},$$

where N_A is the noise added by the amplifier referred to the input.

In radioastronomy and satellite receivers, the generator noise is nowhere near the standard at room temperature, because the noise comes from an antenna that is looking into the sky with a very low effective temperature of only 3 K. In such cases, it does not make sense to talk about a noise figure based on room temperature resistor noise, but rather about amplifier noise temperature T_A given by:

$$T_A = \frac{N_a^o}{k\,G} = \frac{N_A}{k} \tag{87}$$

Now the noise factor (and noise figure) can be related to the noise temperature of the amplifier as:

$$F = 1 + \frac{T_A}{T_o} \tag{88}$$

The noise temperature of a 2-dB noise figure LNA is 170 K, and the best radioastronomy receivers have a noise temperature of 2 K.

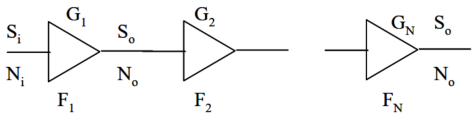


Fig. 39 Noise figure and temperature of an amplifier cascade.

A radio receiver usually receives very low power levels, so more than one amplification stage is needed. Consider a cascade of N amplifiers with respective gains

$G_1, G_2, ...G_N$, as in Fig. 39. What is the equivalent noise figure of this cascade? With all the amplifier noises referred to the inputs of the individual amplifiers, the following follows from the definition of noise figure:

$$F = 1 + \frac{1}{N_i}\left[N_{A1} + \frac{N_{A2}}{G_1} + \frac{N_{A3}}{G_1 G_2} + ... + \frac{N_{AN}}{G_1 G_2 ... G_{N-1}}\right]$$

This is usually written in the following form:

$$F = F_1 + \frac{F_2 - 1}{G_1} + \frac{F_3 - 1}{G_1 G_2} + ... + \frac{F_N - 1}{G_1 G_2 ... G_{N-1}} \tag{89}$$

In the previous formula, the G_i's are available power gains of the individual amplifiers. Therefore, the first amplifier in the chain has the most effect on the overall noise figure, and the cascade design should minimize the first amplifier's noise figure and maximize its gain. Usually stability is an issue with such a cascade and limits the gain. If the chain is connected to a receiving antenna, the above formula tells us that the first LNA should be as close to the antenna feed points as possible. Any length of lossy cable will have negative gain (loss) and its resistance will be a source of noise, so the noise figure of the cascade will be governed by the cable according to the cascade equation.

Given a transistor, how does one design a low-noise amplifier optimized to have the lowest possible noise figure and largest possible gain? The specifications given by the manufacturer give three noise parameters, usually for a number of frequencies: the minimum noise figure (NF_{min}), the optimum input refection coefficient (usually written as Γ_{opt}) and the noise resistance (R_n). These are measured using source-pull tuners (at the input) and a noise-figure meter, and the optimal input impedance is found when the noise figure is measured to be minimal. Then the input of the LNA can be designed by designing a matching circuit for that impedance, while conjugate-matching the output to maximize gain. These parameters, however, do not give us any insight into the source of noise in the transistor, or why there is an optimal input impedance. Furthermore, for very low noise amplifiers, the source-pull method is difficult, since the tuners have to be very stable and calibrated carefully all the time.

To understand where $NF_{min}, \Gamma_{opt}, R_n$ come from, we consider the noise emanating from a transistor described by input and output noise wave quantities and their correlation.

In the case of an amplifier presented as a two-port network (usually, common source), the noise equivalent circuit is given in Fig. 40. The transistor is producing noise waves at both input and output, c_1 and c_2. What are noise waves? Let us start with a simpler, one-port network. A noise wave can be defined for a one-port network as follows. The amplitude of a noise wave c is a random quantity with zero average value, and the expected value $|c|^2$ is noise power, a measurable quantity. When the noise source impedance is equal to the normalizing impedance, the noise power is equal to the power delivered by the source, kT. When the source impedance is not the same as the normalizing impedance, the reflection coefficient of the source is not zero, and in this case the noise wave power per 1Hz bandwidth is equal to

$$\overline{|c|^2} = k\,T\,(1 - |s|^2). \tag{90}$$

This follows from thermal equilibrium. When the source and load are at the same temperatures, if they are matched, the power flow on the line needs to be balanced. If they are not matched, it still needs to remain balanced, otherwise more power would flow in one direction, violating thermal equilibrium. For a one-port network, one can write the following expression for the noise waves in terms of the one-port scattering parameter s:

$$\begin{aligned} b &= s\,a + c \\ \overline{|b|^2} &= |s|^2\,\overline{|a|^2} + \overline{|c|^2} \end{aligned}$$

The latter is true because the incident wave a emanates from the termination and is uncorrelated with the noise wave c. Since thermal equilibrium requires $|b|^2 = |a|^2 = kT$, it follows that:

$$\overline{|c|^2} = kT\big(1 - |s|^2\big) \tag{91}$$

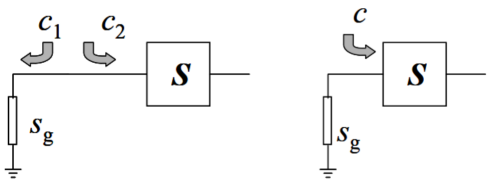


Fig. 40 Equivalent two-port noise waves for an amplifier (a) and the same noise waves referred to the generator side (b).

In Fig. 40 which represents the input of an amplifier, the generator is shown as a one-port with a reflection coefficient s_g. The noise waves can both be referred to the generator in such a way that the total noise at the output of the amplifier remains the same. The new combined noise wave is equal to

$$c = c_1 s_g + c_2.$$

To find the amplifier noise temperature and noise figure, we solve

$$\overline{|c|^2} = kT_A(1 - |s_g|^2) \quad \Rightarrow \quad T_A = \frac{\overline{|c_1 s_g + c_2|^2}}{k(1 - |s_g|^2)}$$

From the relationship between noise factor and noise temperature, the amplifier noise factor is given by

$$F_A = 1 + \frac{\overline{|c_1 s_g + c_2|^2}}{kT_0(1 - |s_g|^2)}.$$

The noise figure can be written as some minimum achievable noise figure, for an optimal input noise match, plus some additional term that depends on the generator reflection coefficient and an optimum reflection coefficient of the generator (s_o):

$$1 + \frac{\overline{|c_1 s_g + c_2|^2}}{kT_0(1 - |s_g|^2)} = F_{\min} + f\frac{|s_g - s_o|^2}{1 - |s_g|^2}.$$

If $\overline{|c_1|^2}$ and $\overline{|c_2|^2}$, as well as their correlation $\overline{c_1 c_2^*}$, are normalized to kT_0, this factor disappears in the denominator of the lefthand side, and the previous equation can be written as

$$\overline{|c_2|^2} + |s_g|^2\overline{|c_1|^2} + 2\Re\left\{s_g^*\,\overline{c_2\,c_1^*}\right\} = (F_{\min}-1)(1-|s_g|^2) + f|s_g|^2 + f|s_o|^2 - 2f\cdot\Re\{s_g^* s_o\}$$

If this is to hold for any generator reflection coefficient, the coefficients of all the powers of s_g must be equal, resulting in the following three equations:

$$\begin{aligned}
\overline{|c_2|^2} &= F_{\min} - 1 + f|s_g|^2 \\
\overline{|c_1|^2} &= -(F_{\min} - 1) + f \\
\overline{c_2 c_1^*} &= -f s_o
\end{aligned}$$

These equations can now be solved to find

$$\begin{aligned}
f &= \frac{\overline{|c_1|^2 + |c_{21}|^2}}{2} + \sqrt{\left[\frac{\overline{|c_1|^2 + |c_{21}|^2}}{2}\right]^2 - |c_1 c_2^*|^2} \\
s_o &= (\Gamma_{\mathbf{opt}} =) - \frac{c_1 c_2^*}{f} \\
F_{\min} &= 1 + f - \overline{|c_1|^2}
\end{aligned}$$

f is found by solving a quadratic equation, and one root of the equation is chosen. The criterion for choosing the root with the plus sign is that the optimum reflection coefficient needs to be smaller than unity. In the process of this derivation, we use the fact that the expected value of the noise power which is a sum of two noise waves is

$$p = \overline{|c_1 + c_2|^2} = \overline{|c_1|^2} + \overline{|c_2|^2} + \overline{c_1 c_2^*} + \overline{c_2 c_1^*}$$

The first two terms are of the noise waves and the last two terms are correlation terms. These two terms are complex conjugates of each other, so their sum is real. The correlation terms are zero if the two noise waves come from physically different noise sources, such as two resistors. If the noises are correlated, such as in a MESFET/HEMT, one tries to choose the reflection coefficient of the generator to cancel as much of the noise as possible at the output. If the two noise waves come from

the same source, they are proportional to each other and they are said to be completely correlated. In this case, the cross-correlation becomes $\overline{|c_1 c_2^*|^2} = \overline{|c_1|^2 |c_2|^2}$.

By observing the expressions for the optimal reflection coefficient, it is seen that it is proportional to the correlation of the two noise waves. If the two noise sources are uncorrelated, then $s_o = 0$, but in a MESFET/HEMT usually the magnitude of the optimum generator scattering coefficient is close to unity, and this indicates that the two noise sources are closely correlated and that they physically come from one source inside the transistor.

In transistor specification sheets, usually a minimum noise figure and optimal scattering parameter are given for a number of frequencies. Instead of f, a related parameter referred to as the normalized noise resistance is given:

$$\frac{f}{k \cdot 290\,\mathrm{K}} = \frac{4R_n}{Z_0|1 + s_o|^2}$$

What are the sources of noise in a MESFET/HEMT? With reference to a very simplified unilateral mode, the resistance between the gate and source produces a noise voltage that becomes a part of the input voltage, and is coupled to the output through the transconductance. It can be shown that the input and output noise power spectral densities and the input and output noise wave correlation for a transistor can be writing in terms of the device *S*-parameters and the effective noise temperatures of the gate and drain (derived from Bosma's theorem). [A nice paper that describes this is: S. W. Wedge and D. B. Rutledge, "Wave techniques for noise modeling and measurement," in *IEEE Transactions on Microwave Theory and Techniques, vol.* 40*,* no. 11, pp. 2004-2012, Nov. 1992.] Since the transistor is not passive, thermal equilibrium can be achieved only if the dc input is modeled by an effective increase in noise temperature. In practice, the drain effective noise temperature is the main effect.

7.4 Noise Measurements

For a MESFET or HEMT, as mentioned, a minimum noise figure and corresponding optimal input reflection coefficient are usually given. These are measured using a tuner at the input of the device (referred to as a"source tuner"), while simultaneously measuring the noise figure. How is the noise figure measured?

There are several methods for determining the noise figure of an amplifier. A method easy to understand, but not so easy to implement accurately in practice, is the Y-parameter measurement, or hot-and-cold measurement, Fig. 41. First a resistor R connected as the noise source to the input of an amplifier is heated to a temperature T_H and then cooled to a temperature T_C. The resulting measured noise power densities in a bandwidth Δf at the output of the amplifier are

$$P_H = G\,k\,T_H\,\Delta f + G\,k\,T_A\,\Delta f$$
$$P_C = G\,k\,T_C\,\Delta f + G\,k\,T_A\,\Delta f$$

The Y-factor is defined as the ratio of these two noise powers and can be written as

$$Y = \frac{T_A + T_H}{T_A + T_C} \quad \text{so that} \quad T_A = \frac{T_H - Y\,T_C}{Y - 1}$$

By measuring the Y factor, the noise temperature of the amplifier, and therefore the noise figure, are determined. The accuracy of the measurement depends on the two temperature being as far apart as possible, and requires that the reflection coefficient of the load resistor R does not vary with temperature. Since we start with two equations with two unknowns (the gain is also unknown), the gain of the amplifier can also be determined. Since the gain can be measured independently with a network analyzer, this can help check or improve the accuracy of the measurement.

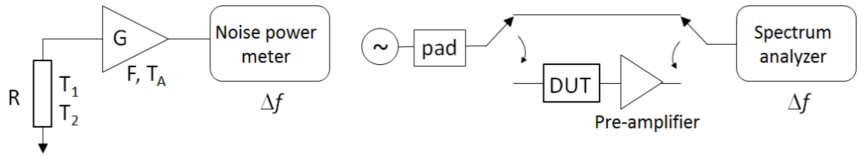


Fig. 41 (a) Hot and cold (Y-factor) noise figure measurement, (b) spectrum analyzer noise figure measurement and (c) automated noise figure meter block diagram.

Another way to measure noise figure is directly with a spectrum analyzer, Fig. 42. The method is not very accurate for noise figures below 3 dB. A known pre-amplifier, with gain $G_{pre-amp}$ and output noise spectral power density $N_{pre-amp}$ is used in one branch of the system. For a limited frequency bandwidth, the spectrum analyzer IF bandwidth is set to B, and the resulted measured DUT noise figure is

$$F = \frac{N_{\text{pre-amp}}}{k\,T\,G_{\text{pre-amp}}\,G_A\,B}$$

where G_A is the unknown amplifier's (DUT's) gain. The noise power measured by the spectrum analyzer integrated over the IF bandwidth will have a correction constant of C=1-2 dB which includes different detection effects, for example the video filter used for the post-detection display, which is usually set to one hundredth of the IF bandwidth. In dB, the noise figure can be written in terms of the measured power by the spectrum analyzer:

$$NF\text{dB} = P_n - G_A - G_{\text{pre-amp}} - 10\log B + 174 + C$$

(Where does the 174 number come from?). In the automatic noise figure meter, a noise source is used to provide different levels of noise. This is usually a pn diode

biased in reverse avalanche breakdown, and the available noise power is inversely proportional to the diode current.

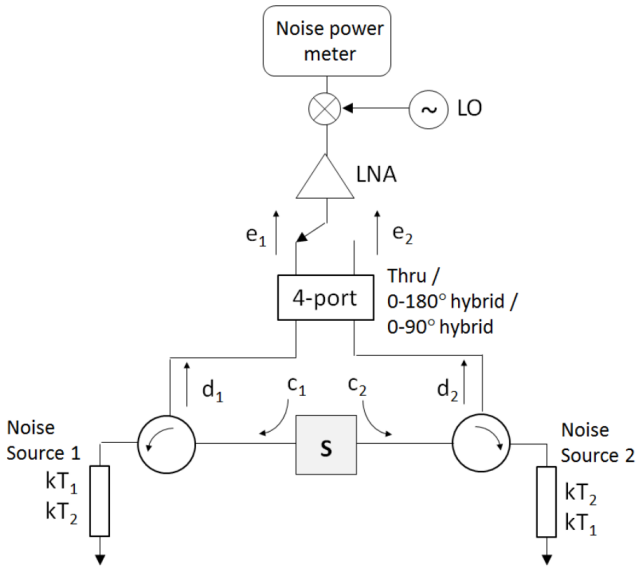


Fig. 42 Noise wave measurement block diagram.

Noise figure can also be measured indirectly by measuring noise waves. This is especially useful for characterizing transistors for the optimal input noise match. The measurement does not require a source tuner, and eliminates errors due to tuner limited range and tuner calibration. Consider the system in Fig. 42. The noise waves emanating from the device are c_1 and c_2. They are added to noise waves generated by two noise sources that can each be at two equivalent temperatures, T_1 and T_2. The resulting wave power spectral densities at the input of the switched four-port network are:

$$\begin{aligned}
\overline{|d_1|^2} &= \overline{|c_1|^2} + k\,T_1|s_{11}|^2 + k\,T_2|s_{12}|^2 \\
\overline{|d_2|^2} &= \overline{|c_2|^2} + k\,T_1|s_{21}|^2 + k\,T_2|s_{22}|^2 \\
\overline{d_1 d_2^*} &= \overline{c_1 c_2^*} + k\,T_1 s_{11} s_{21}^* + k\,T_2 s_{12} s_{22}^*
\end{aligned}$$

These are measurable quantities, and in order to obtain from them the unknown c_1 and c_2, first a "thru" is inserted in the place of the switched four-port. In this case, a noise power measurement is performed at each temperature of the noise source, resulting in four noise power measurements. To measure the correlation, first a 0-180 degree 3-dB coupler, and then a 0-90 degree coupler are inserted in the place of the four-port. The first set of measurements at the two temperatures of the noise sources

gives the real part of the correlation, and the second set gives the imaginary part. For example, for the 0-180 degree hybrid case, the following can be written:

$$e_1 = \frac{1}{\sqrt{2}}(d_1 + d_2)$$
$$e_2 = \frac{1}{\sqrt{2}}(d_1 - d_2)$$
$$\overline{|e_1|^2} - \overline{|e_2|^2} = 2\Re\left(\overline{d_1 d_2^*}\right) = 2\left[\Re\left(\overline{c_1 c_2^*}\right) + k\,T_1(\overline{s_{11}s_{21}^*}) + k\,T_2(\overline{s_{12}s_{22}^*})\right]$$

The resulting number of measurements provide an overdetermined system of equations, and in addition to obtaining the *S*-parameters of the device, some statistical analysis of these measurements can be performed.

8 Power Amplifiers

Power amplifiers are treated nicely in several books, please look at one of these and follow the discussions in the indicated chapters:

Inder J. Bahl,"Fundamentals of RF and Microwave Transistor Amplifiers, " Inder J. Bahl, Wiley 2009, Chapter 8.

Steve Cripps, "RF Power Amplifiers for Wireless Communications," Second Edition, Artech House 2006, Chapters 1, 2 and 3.

8.1 Class-E Switched-Mode Power Amplifiers

In a switched-mode PA, such as class E or D, the transistor is used as a switch driven by the (input) RF signal. One can define a power gain, but it is not quite the same as for the previously discussed classes of operation. In this case, see Fig. 43, the DC supply voltage is switched on and off and the resulting waveform filtered to obtain a sinusoidal current through the load. The biggest hurdle in operating the device as a switch at microwave frequencies is the output capacitance of the transistor, and in the circuit in Fig. 43, it is de-embedded and taken out of the device which is modeled as an ideal switch. the design of a switched-mode PA reduces to the design of the output circuit which ensures that the current and voltage across the device overlap a minimal percentage of the period. The class E of operation is defined by specific waveform shapes in the time domain which ensure high efficiency.

There are a number of assumptions which are made in order to obtain solutions for the class-E circuit, i.e. the elements of the output resonant circuit. First, the switching is assumed to be at a 50% duty cycle. Next, it is assumed that the switching element

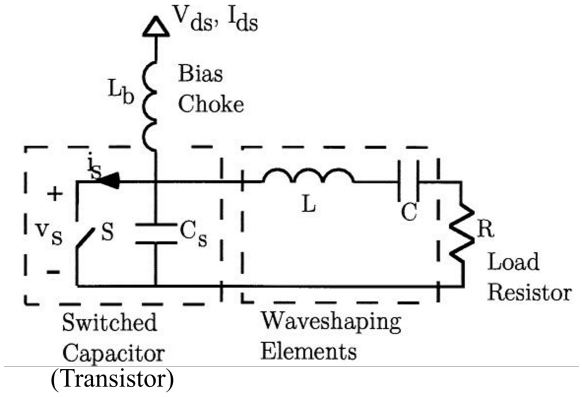


Fig. 43 Original Class-E PA topology as introduced by Sokal and Sokal at low frequencies, where the capacitor C_S had to be added externally and lumped elements behave as almost ideal.

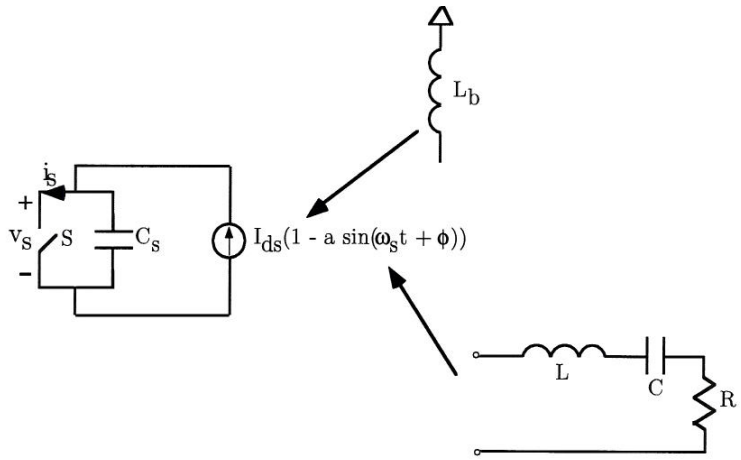


Fig. 44 Class-E PA approximate circuit when the transistor is assumed to be a switch with a linear parallel capacitance, and the resonator is replaced with a Norton equivalent sinusoidal current source, with the DC current through an ideal source included.

has zero on-resistance and infinite off-resistance (although these can later be taken into account). The output capacitance of the device is assumed to be linear (not dependent on voltage). The bias choke is assumed to be ideal (infinite impedance at RF). It is also assumed that the elements of the circuit are lossless so that all of the input power $I_{ds}V_{ds}$ is delivered to the load (100% efficiency). The external load LCR network has a high Q factor and is resonant slightly below the switching frequency, so that the voltage through the load is sinusoidal, and the current through the inductor is sinusoidal at the switching frequency.

Fig. 44 shows a simple Norton circuit that can be used to analyze a class-E PA. The fact that under ideal conditions the current through the load is sinusoidal is used to replace the load by a Norton equivalent current source with, so far, unknown amplitude and phase at the switching frequency. When the switch is ON, there is no voltage across the switch, but a part of the sinusoidal current with a DC component flows through it. At the instant the switch is turned on, the current through the switch is zero, at when it is turned off, there is a discontinuous jump in current. This jump in current will cause losses to appear across any parasitic inductance between the transistor and the shunt capacitor with an energy loss of $LI^2/2$ per period.

When the switch is OFF, the sinusoidal current continues to flow, now through the shunt capacitor of the switch. During the off interval, we can write the following equation and boundary conditions:

$$C_s \frac{\mathrm{d}v_s}{\mathrm{d}t} = I_{ds}\,(1 - a\sin(\omega_s t + \phi) \tag{92}$$

$$v_s(t) = \frac{I_{ds}}{C_s}\int_0^t (1 - a\sin(\omega_s t' + \phi))\mathrm{d}t' \tag{93}$$

$$v_s(t) = \frac{I_{ds}}{\omega_s C_s}(\omega_s t + a(\cos(\omega_s t + \phi) - \cos\phi)) \tag{94}$$

$$v_{Cs}(t=0) = 0, \qquad v_s\left(\frac{T_s}{2}\right) = 0 \qquad \text{and} \qquad \frac{\mathrm{d}v_s}{\mathrm{d}t}\left(\frac{T_s}{2}\right) = 0 \tag{95}$$

The first boundary condition assumes that there is no voltage across the capacitor at time zero, the second avoids shorting the capacitor when there is a voltage across it, and the third ensures a "soft" turn-on condition for the device. This is important, as it makes the circuit implemented based on this analysis less sensitive to changes in element values. Now we can determine a and ϕ uniquely as follows:

$$a = \sqrt{1 + \frac{\pi^2}{4}} \approx 1.862, \tag{96}$$

$$\phi = -\arctan\frac{2}{\pi} \approx -32.48^\circ. \tag{97}$$

Note that these are constants for any high-Q class-E circuit with a capacitor in shunt with the switch. The voltage and current across the switch are now known and have the following form, shown in Fig. **??**:

$$v_s(t) = \frac{I_{ds}}{\omega_s C_s}((\omega_s t) + a(\cos((\omega_s t) + \phi - \cos\phi)), \quad \text{for} \quad 0 \le (\omega_s t) \le \pi \tag{98}$$

$$v_s(t) = 0, \quad \text{for} \quad \pi \le (\omega_s t) \le 2\pi \tag{99}$$

$$i_s(t) = 0, \quad \text{for} \quad 0 \le (\omega_s t) \le \pi \tag{100}$$

$$i_s(t) = I_{ds}(1 - a\sin(\omega_s t + \phi)), \quad \text{for} \quad \pi \le (\omega_s t) \le 2\pi. \tag{101}$$

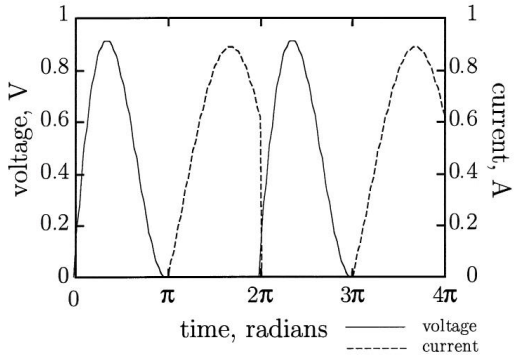


Fig. 45 Ideal class-E time-domain waveforms for the voltage and current across the switching device.

It is important to find the relationship between the drain voltage and current, i.e. how much current is drawn for a given supply voltage. V_{ds} is the DC component of $v_s(t)$. Take the time average of the switch voltage to obtain:

$$V_{ds} = \frac{1}{T_s}\int_0^{\frac{T_s}{2}} v_s(t)\mathrm{d}t = \frac{1}{\pi}\frac{I_{ds}}{\omega_s C_s}, \tag{102}$$

$$I_{ds} = \pi\omega_s C_s V_{ds} \tag{103}$$

This simple result has important implications for a practical microwave class-E PA. If Cs is the output capacitance of the device, it means that at a specified frequency, a device with a given output capacitance must operate at some supply voltage, well above the knee. Then the device must be able to handle the maximal current given below. Alternatively, one can set the maximal current and then the voltage is a few times the supply voltage:

$$i_{max} = (1 + a) \cdot I_{ds} = 2.6 \cdot I_{ds} \quad \text{and} \quad v_{max} \approx 3.56 \cdot V_{ds}. \tag{104}$$

Looking at the relationship in a different way, for a given voltage, current and capacitance, there is a certain highest frequency at which a device can operate in ideal class-E mode:

$$f_{max,E} = \frac{I_{ds}}{2\pi^2 C_s V_{ds}} = \frac{I_{max}}{C_s V_{ds}}\frac{1}{2\pi^2(1 + a)} \approx \frac{I_{max}}{56.5\, C_s V_{ds}} \tag{105}$$

This is very important practically, as you can determine, given device capacitance (size) and current/voltage handling, how high in frequency an efficient class-E PA can be designed with it. Fortunately, if one wants to use a device at higher frequencies than $f_{max,E}$, it is still possible to use these design equations and the PA will gracefully degrade into a still efficient class-AB PA. To find the load network required for class-E waveforms, we can expand the voltage in Fourier components and find the impedance at the fundamental:

$$v_s(t) = \sum_{n=-\infty}^{\infty} K_n e^{jnw_s t}, \tag{106}$$

$$K_n = \frac{1}{T_s} \int_0^{\frac{T_s}{2}} v_s(t) e^{-jnw_s t} \mathrm{d}t \tag{107}$$

$$K_1 = \frac{I_{ds}}{\omega_s C_s T_s} \int_0^{\frac{T_s}{2}} (\omega_s t + a(\cos(\omega_s t + \phi) - \cos\phi)) e^{-j\omega_s t} \mathrm{d}t \tag{108}$$

Solving the above equations, which is lengthy but straightforward, one obtains:

$$v_{s1} = a_0 I_{ds} \sin(\omega_s t + \phi_0) \tag{109}$$

$$i_{net_1} = a I_{ds} \sin(\omega_s t + \phi) \tag{110}$$

$$a_0 = \frac{2|K_1|}{I_{ds}} = \frac{1}{\omega_s C_s} \sqrt{\frac{\pi^2}{16} + \frac{4}{\pi^2} - \frac{3}{4}} \quad \text{and} \quad \phi_0 = \frac{\pi}{2} + \angle K_1 = \frac{\pi}{2} + \arctan\left(\frac{2\pi}{8 - \pi^2}\right) \tag{111}$$

Now dividing voltage by current, we obtain an expression for the impedance at the switching frequency, and this is independent of device, circuit topology etc.:

$$Z_{net_1} = \frac{a_0}{a} e^{j(\phi_0 - \phi)} \approx \frac{0.28015}{\omega_s C_s} e^{j49.0524^\circ} \tag{112}$$

In terms of the maximum ideal class-E frequency of operation, as an example, for a general purpose FET, with $I_{max} = 1200\,mA$, $C_s \approx 2.6\,\mathrm{pF}$, $V_{ds} = 6\,\mathrm{V}$, the maximal frequency of operation in ideal class-E mode is $f_{max,E} \approx 1.4\,\mathrm{GHz}$. If the drain voltage is reduced this frequency increases and power decreases. The rule of thumb is that this is the price to pay for high efficiency without sacrificing more than 1 dB of output power: the device has to be able to operate at 3-5 times higher frequency than the switching frequency.

Now the specific load topology needs to be determined. The series RLC ends up not being a good topology (over-constrained), but the one shown in Fig. 46 is convenient, giving:

$$Z_{net_1} = j\omega_s L + \frac{1}{j\omega_s C} + R$$

This circuit can also be implemented with transmission lines (series line for L and shunt open stub for C). This transmission-line equivalent circuit is shown in Fig. 47.

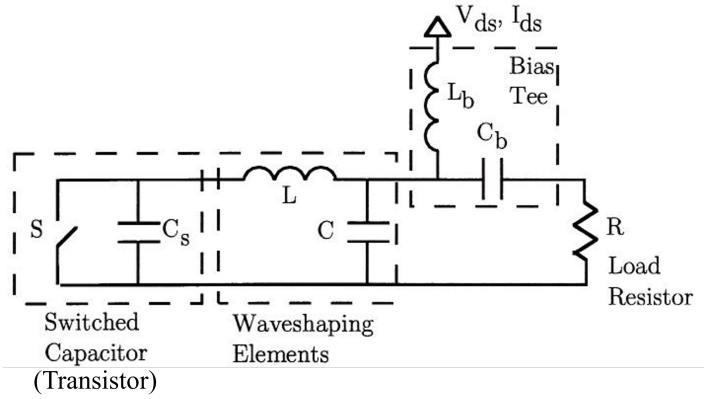


Fig. 46 Lumped-element load circuit for ideal class-E operation when device is modeled as an ideal switch in parallel with a capacitance.

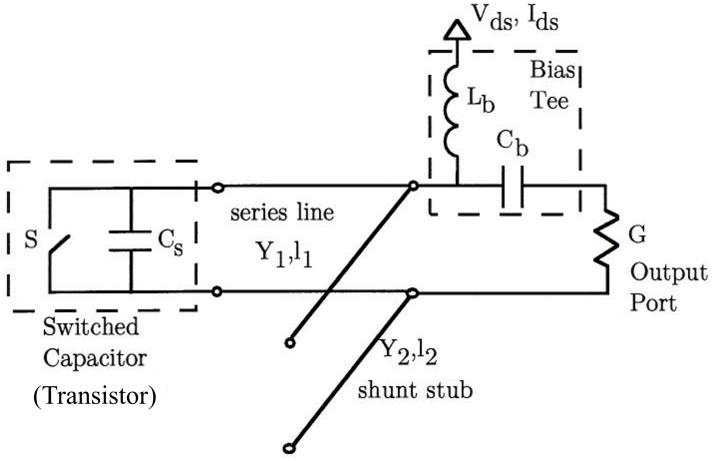


Fig. 47 Transmission-line load circuit for ideal class-E operation when device is modeled as an ideal switch in parallel with a capacitance.

The elements of the transmission line circuit can easily be found numerically and using standard transmission line equations. Note that the class-E topology assumes that the harmonics are terminated in opens. This can be enforced by adding stubs that open the harmonic, but are a part of the fundamental class-E impedance.

It is off course impossible to get a 100% efficient PA, and there are losses that will limit this to demonstrated > 80% in the lower microwave frequency range and > 70% even as high as Ka-band. The efficiency is limited by the off and on resistances of the switch, and these can be taken into account by re-deriving the above equations with a resistance in parallel with the switch capacitance, as sketched in Figure L10.9.

If one goes through the equations in the same manner as for the case without the resistance, the efficiency becomes:

$$\eta_D = \frac{1 + (\pi/2 + \omega_s R_s C - S)^2}{(1 + \pi^2/4)(1 + \pi\omega_s R_s C_s)^2}$$

This expression is sketched in Fig. 48 as a function of the product of the $R_s C_s$ time constant and frequency.

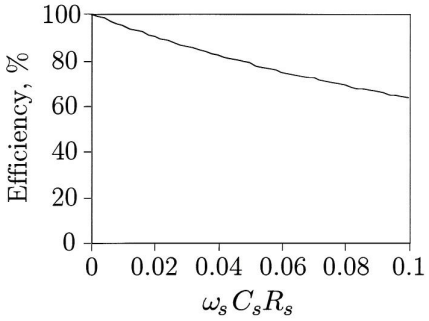


Fig. 48 Load circuit for ideal class-E operation when device is modeled as an ideal switch in parallel with a capacitance.

9 Harmonic Balance Fundamentals

The *S*-parameter analysis we have done so far is based on linear parameters: for a single sinusoidal excitation, the output will contain only the same sinusoidal frequency. Soon we will analyze oscillators, in which the input is a DC voltage, while the desired output is an RF sinusoidal voltage, with a number of harmonic frequency components present. The many frequencies present in the output voltage with a DC input were a result of the nonlinear characteristics of the transistor.

There are two ways to analyze nonlinear circuits: in time domain (such as in Spice) or in frequency domain by separately considering the response of the circuit to a number of harmonics of the fundamental input frequency. Time-domain techniques have the following problems at microwave frequencies:

- it is difficult to include linear matching and other networks that are dispersive (frequency-dependent);

- long integrations accompanied by intensive computations and large truncation errors are necessary when the time-constant of circuit is large compared to the period of the excitation frequency; and
- each linear or nonlinear reactive element in the circuit adds a differential equation to the set of equations that describe the circuit.

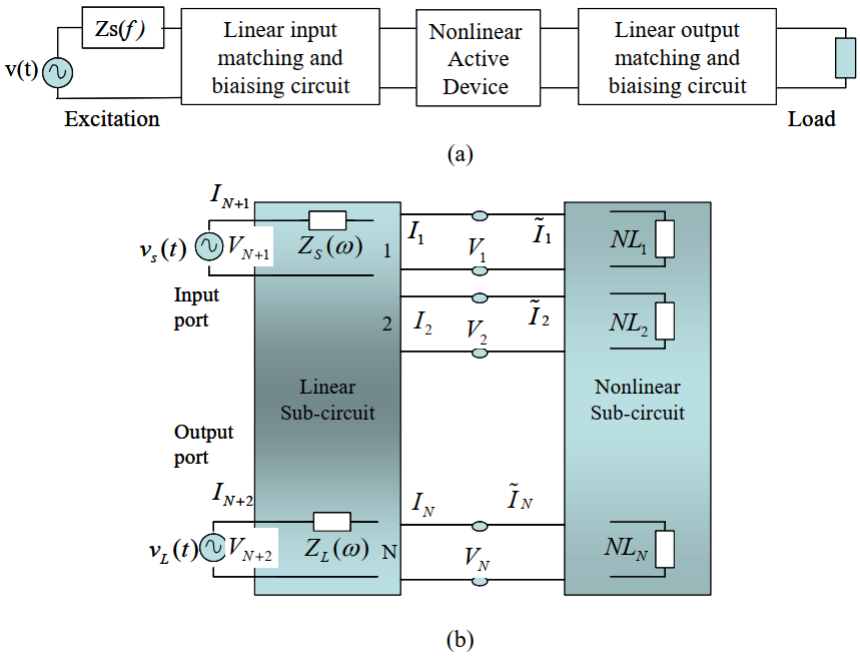


Fig. 49 A nonlinear microwave circuit block diagram (a) and re-grouped elements of the circuit for harmonic balance analysis (b).

Instead, in microwave circuit simulators, multiport theory combined with frequency-domain simulations is used. Consider the circuit in Fig. 49a, which contains some nonlinear active device that requires biasing (a diode or transistor) and linear matching and filtering networks. The circuit elements can be regrouped in such a way that all linear components (including linear parts of the active device) are contained in one multiport network with $N + 2$ ports, while the nonlinear components are grouped on a non-linear multiport network with N ports, Fig. 49b. Each of the N ports contains all frequency components of interest, and the current continuity can be expressed as

$$\begin{bmatrix} I_{1,0} \\ I_{1,1} \\ I_{1,2} \\ \vdots \\ I_{N,K} \end{bmatrix} + \begin{bmatrix} \widetilde{I}_{1,0} \\ \widetilde{I}_{1,1} \\ \widetilde{I}_{1,2} \\ \vdots \\ \widetilde{I}_{N,K} \end{bmatrix} = \underline{\underline{0}}, \tag{113}$$

where the index $1..K$ represents the harmonic of the 1st through N-th port node. The basic principle of harmonic balance is to solve the circuit to find port voltages such that Eq.(KCL-HB) is satisfied, i.e. such that the currents in the nonlinear and linear sub-network connecting ports are identical.

The sources at the input and output ports are given for generality, and their impedances are included in the linear sub-circuit. Usually, a sinusoidal source is present only at the input, while the output port contains only DC bias (such as in a FET amplifier). The voltages and currents at the N ports contain the DC component, as well as K harmonics of the fundamental excitation frequency, and it is assumed that these K harmonics adequately describe the nonlinear circuit. In many cases, the choice of K is critical to accurate modeling.

The linear sub-circuit admittance (Y) matrix is a $(N+2)\times(N+2)$ matrix relating the $(N+2)$ currents at the K harmonics to the $(N+2)$ port voltages at the K harmonics:

$$\begin{bmatrix} \mathbf{I}_1 \\ \vdots \\ \mathbf{I}_{N+2} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{11} & \dots & \mathbf{Y}_{1,N+2} \\ \vdots & \vdots & \vdots \\ \mathbf{Y}_{N+2,1} & \dots & \mathbf{Y}_{N+2,N+2} \end{bmatrix} \begin{bmatrix} \mathbf{V}_1 \\ \vdots \\ \mathbf{V}_{N+2} \end{bmatrix}, \tag{114}$$

where each of the bold-faced current and voltage sub-matrices contain the K harmonics and are given by

$$\mathbf{I_n} = \begin{bmatrix} I_{n0} \\ I_{n1} \\ \vdots \\ I_{nK} \end{bmatrix} \quad \text{and} \quad \mathbf{V_n} = \begin{bmatrix} V_{n0} \\ V_{n1} \\ \vdots \\ V_{nK} \end{bmatrix}$$

The Y submatrices are diagonal matrices with elements given by

$$\mathbf{Y_{m,n}} = \begin{bmatrix} Y_{m,n}(0) & 0 & \dots & 0 \\ 0 & Y_{m,n}(\omega) & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & Y_{m,n}(K\omega) \end{bmatrix}.$$

The two voltage harmonic vectors corresponding to the input and output ports can be written as an “excitation vector”:

$$\begin{bmatrix} \mathbf{V}_{N+1} \\ \mathbf{V}_{N+2} \end{bmatrix} = \begin{bmatrix} V_{b1} \\ V_S \\ 0 \\ \vdots \\ V_{b2} \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

where V_{b1} and V_{b2} are bias voltages at the input and output ports $(N+1)$ and $(N+2)$, and V_s is the input excitation voltage at port (N+1). Port (N+1) has both an AC and DC input, while port (N+2) has only a DC bias, such as in the case of a FET amplifier. If the input is not sinusoidal, the vector on the right would include harmonic voltages instead of 0s. Now the Y matrix in Eq.(114) can be partitioned and expressed in terms of a 2×2 excitation matrix and a $N \times N$ circuit matrix:

$$\begin{bmatrix} \mathbf{I}_1 \\ \vdots \\ \mathbf{I}_N \end{bmatrix}_{N\times 1} = \begin{bmatrix} \mathbf{Y}_{1,N+1} & \mathbf{Y}_{1,N+2} \\ \vdots & \vdots \\ \mathbf{Y}_{N,N+1} & \mathbf{Y}_{N,N+2} \end{bmatrix}_{N\times 2} \begin{bmatrix} \mathbf{V}_{N+1} \\ \mathbf{V}_{N+2} \end{bmatrix}_{2\times 1} + \begin{bmatrix} \mathbf{Y}_{1,1} & \mathbf{Y}_{1,N} \\ \vdots & \vdots \\ \mathbf{Y}_{N,1} & \mathbf{Y}_{N,N} \end{bmatrix}_{N\times N} \begin{bmatrix} \mathbf{V}_1 \\ \vdots \\ \mathbf{V}_N \end{bmatrix}_{N\times 1}$$

The above matrix equation can be re-written as $\mathbf{I} = \mathbf{I}_S + \mathbf{Y}_{NxN} \cdot \mathbf{V}$. The circuit-theory meaning of this equation is as follows. $\mathbf{I}_S$ represents N current sources in parallel with the N ports. The first term in the matrix equation transforms the excitations at the input and output ports into these parallel current sources, as shown in Fig. 50. This eliminates the need to keep track of the $(N+1)$st and $(N+2)$nd ports. Therefore, the harmonic-balance equations are expressed in terms of only the currents in the N ports connected to the nonlinear elements. The currents in the nonlinear elements, $\tilde{I}$, can be a result of non-linear capacitors or non-linear resistors.

Now what remains to be answered is how the different nonlinear elements are included in the analysis. In a microwave diode, there are two nonlinearities: the resistance (slope of the IV curve) and the junction capacitance. In a transistor, there are a number of nonlinear elements, for example the voltage-dependent current source and the output capacitance are nonlinear elements. A voltage-controlled current source can be described by the very general time-domain equation:

$$i_{g,n}(t) = f_n(v_1(t), v_2(t), ..., v_n(t))$$

which can be Fourier-transformed to obtain an expression in frequency domain:

$$\mathbf{F}\{i_{g,n}(t)\} \rightarrow \mathbf{I}_{G,n}.$$

A vector of all $n = N$ currents is the input $\mathbf{I}_G$ to our harmonic balance equation.

In a nonlinear capacitor, the charge can be expressed in terms of voltages in time domain, and then a Fourier transform performed:

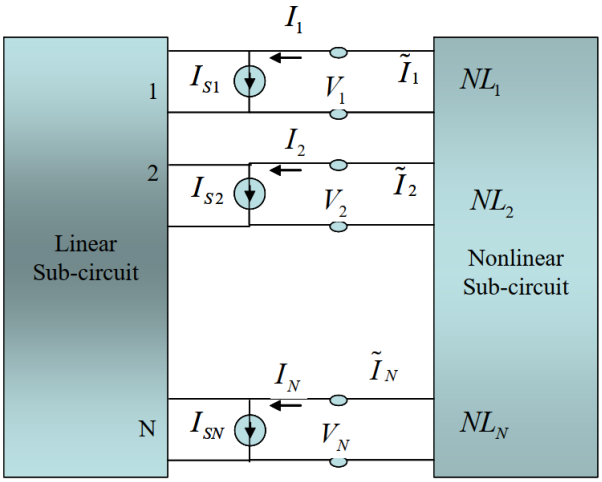


Fig. 50 Re-grouped equivalent circuit, with the excitation (AC and DC) signals represented as equivalent current sources distributed in the N ports.

$$q_n(t) = f_{qn}(v_1(t), v_2(t), ..., v_n(t)) \quad \text{and} \quad \mathrm{F}\{q_n(t)\} \rightarrow \mathbf{Q}_n$$

where the charge vector input to the harmonic balance matrix equation is

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \\ \vdots \\ \mathbf{Q}_N \end{bmatrix} = \begin{bmatrix} Q_{1,0} \\ Q_{1,1} \\ Q_{1,2} \\ \vdots \\ Q_{1,K} \\ Q_{2,0} \\ \vdots \\ Q_{2,K} \\ \vdots \\ Q_{N,K} \end{bmatrix},$$

and where N is the number of nonlinear capacitors and K is the number of harmonic frequencies. The current in the nonlinear capacitor is the time derivative of the charge waveform, which corresponds to a multiplication with $j\omega$ in the time-harmonic domain:

$$i_{c,n}(t) = \frac{\mathrm{d}q_n(t)}{\mathrm{d}t} \Leftrightarrow jk\omega Q_{n,k},$$

which can be expressed in matrix form by introducing a "harmonic frequency" matrix which is diagonal and has N cycles of $[0, \omega, 2\omega, ..., K\omega]$ along the diagonal,

and is a large $(K+1)N \times (K+1)N$ matrix. The capacitor current vector input to the harmonic balance matrix equation can now be written as

$$\mathbf{I}_C = j\mathbf{\Omega Q},$$

where $\mathbf{\Omega}$ is a diagonal matrix of frequencies $\omega, 2\omega, ..., K\omega$. The nonlinear capacitor and nonlinear resistor/conductor currents are now included in Eq. 113 as follows:

$$\mathbf{F(V)} = \mathbf{I}_S + \mathbf{Y}_{N\times N}\mathbf{V} + j\mathbf{\Omega Q} + \mathbf{I}_G = \mathbf{0}. \tag{115}$$

This equation is a test to determine the accuracy of the current continuity at the nodes of the circuit separated into a linear and nonlinear subcircuit, and assuming a trial set of voltages at the ports. If Eq. 115 holds, the voltages of vector $\mathbf{V}$ are the solutions. $\mathbf{F}(\mathbf{V})$ is the current error vector of the harmonic balance equation 115.

Each of the $(K+1)$ frequency components of $\mathbf{V}$ at each port is a variable, with real and imaginary (or phase and amplitude) parts, resulting in $2N(K+1)$ unknowns to solve for. As an example, in a FET multiplier, there are 3 nonlinear ports, and usually no less than 8 frequencies are needed for a reasonably accurate solution (assuming a good transistor model). Therefore, in the multiplier harmonic balance solution, there are 54 unknown voltages to solve for. What numerical methods are used to obtain these solutions?

A number of numerical methods seem like a reasonable approach, for example, optimization, Newton's method (using the Jacobian of $\mathbf{F}(\mathbf{V})$), and the more physically intuitive splitting and reflection methods. The reflection method was introduced by Kerr to study harmonically-pumped mixers, and this will be a topic of discussion in one of the next lectures.

The splitting method starts with an estimate $\mathbf{V}^0$ of the solution. Then $\mathbf{V}^0$ is Fourier-transformed to obtain $v_n^0(t)$, and $v_n^0(t)$ is substituted into the nonlinear element equations to obtain current and charge waveforms. These are then Fourier transformed. $\mathbf{I}^0$ is then estimated as $\tilde{\mathbf{I}}^0$ and a new voltage vector is found form the linear subcircuit:

$$\mathbf{V}'' = \mathbf{Y}_{N\times N}^{-1}(\mathbf{I}^0 - \mathbf{I}_S)$$

The new estimate of $\mathbf{V}$ is then found: $\mathbf{V}^1 = s\mathbf{V}'' + (1-s)\mathbf{V}^0$. Here, s is a real number referred to as the splitting coefficient and is a constant that is determined empirically, but is usually 0.2 and never bigger than 1 (small values yield slow convergence, while larger ones can give instable solutions). Next, the process is repeated with $\mathbf{V}^0$ replaced by $\mathbf{V}^1$, and so on until a pre-specified satisfactory value of the error.

9.1 Harmonic Balance Example

As a simple illustration of the harmonic balance method, consider a detector circuit (we will study detectors shortly), as shown in Fig. 51a. The circuit is rearranged

into linear and nonlinear subnetworks and the excitation, Fig. 51b. Subsequently, following the discussion above, the excitation is represented as a current source effectively transforming the circuit to an equivalent one-port. We will assume that the diode is ideal and has no series resistance or junction capacitance. Therefore, only one nonlinear element exists, so $N = 1$, and the vector $\mathbf{V} = \mathbf{V_1}$. The admittance matrix of the linear embedding network, $\mathbf{Y}_m$, can be written as:

$$\mathbf{Y}_m = \begin{bmatrix} \mathbf{Y}_{1,1} & \mathbf{Y}_{1,2} \\ \mathbf{Y}_{2,1} & \mathbf{Y}_{2,2} \end{bmatrix}, \text{ and } \mathbf{I}_S = \mathbf{I}_{S,1} = \mathbf{Y}_{1,2}\mathbf{V}_2.$$

When $\mathbf{V}_2$ is transformed through the Y network, the equivalent circuit from Fig. 51c results. The vector $\mathbf{V}_2$ consists only of the fundamental frequency source $V \cos \omega t$ and the DC bias source:

$$\mathbf{V}_2 = \begin{bmatrix} V_b \\ V \\ 0 \\ \vdots \end{bmatrix}$$

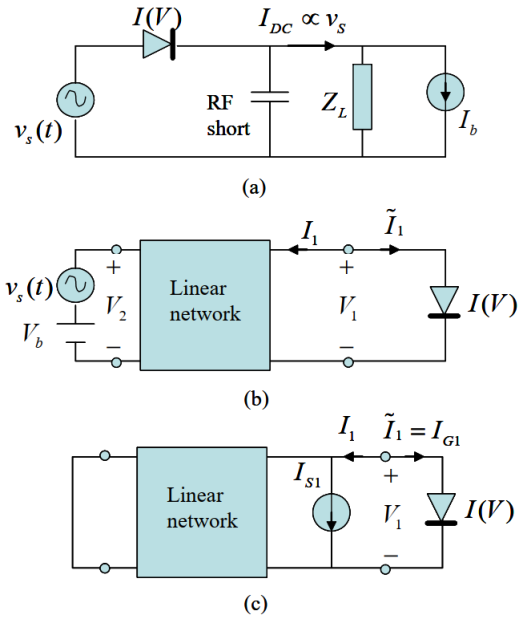


Fig. 51 (a) Video detector circuit. (b) Rearranged circuit with linear and nonlinear subnetworks, and (c) with two-port equivalent circuit reduced to one-port.

Now an initial estimate of either $v_1(t)$ or $\mathbf{V}_1$ is needed, and usually for a diode it is taken as a clipped sinusoid (clipped at about 0.6 V). The Fourier transform of $v(t)$

gives the components of $\mathbf{V}_1$. The diode current is given by the ideal diode equation:

$$i_{g,1}(t) = I_s(\exp(v_1(t)/V_T) - 1) \text{ and since there is no capacitance, } q(t) = 0.$$

and since there is no capacitance, $q(t) = 0$. The current error vector is now:

$$\mathbf{F}(\mathbf{V}) = \mathbf{Y}_{1,2}\mathbf{V}_2 + \mathbf{Y}_{1,1}\mathbf{V}_1 + \mathbf{I}_{G,1}.$$

In order to solve this equation, the splitting algorithm can be used as follows:

- Invert $\mathbf{Y}_{1,1}$ matrix (diagonal) to generate embedding impedance matrix;
- form an initial estimate of $v_1(t)$ as a sinusoid at the fundamental frequency clipped by 0.6V with a DC offset of the bias voltage;
- calculate $i_{g,1}(t) = I_s(\exp(v_1(t)/V_T) - 1)$;
- Perform Fourier transforms of $i_{g,1}(t)$ and $v_1(t)$, which gives $\mathbf{I}_G^0$ and $\mathbf{V}_1^0$;
- Assume $\mathbf{I} = -\mathbf{I}_{G,1}^0$ and form $\mathbf{V}^n = \mathbf{Z}_{1,1}(\mathbf{I} - \mathbf{I}_s)$;
- Form the new estimate of $\mathbf{V}_1$ as $\mathbf{V}_1^1 = s\mathbf{V}'' + (1 - s)\mathbf{V}_1^0$;
- Perform inverse Fourier transform of $\mathbf{V}_1^1$ to obtain a new estimate of $v_1(t)$, repeat step 3 until convergence is achieved;
- If there is an instability, reduce s.

The two other Y parameters of the admittance matrix that were not mentioned above can be used to find the input current from the source ($\mathbf{I}_2$) once $\mathbf{V}_1$ is known. $\mathbf{I}_2$ is then used to find the input power, the power dissipated in the source and the source impedance that was previously included in the linear network.

10 Microwave Oscillators

A microwave oscillator is a circuit that converts DC power to microwave power. The standard feedback model of an oscillator is shown in Fig. 52. There is a gain element (transistor) and a passive circuit, and the passive circuit is a resonator of some type. The phase of the gain (loss) is a function of frequency, and the amplitude of the gain (loss) is a function of power as shown qualitatively in the figure. The oscillation condition is met when $|G| = |L|$ and $\angle G = \angle L$. We see that the frequency of steady-state oscillation, when $\angle G = \angle L(f_0)$, will be largely determined by the passive resonant circuit since its phase changes very fast with frequency. On the other hand, the power will be determined by the active gain element when $|G(P0)| = |L|$. Therefore, in an oscillator design, the frequency of operation is determined by the design of the resonant circuit, which can have many forms at microwave frequencies as described below.

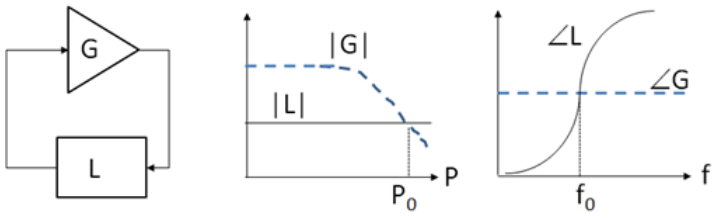


Fig. 52 (a) Basic block diagram of an oscillator and qualitative dependence of gain and loss on frequency and power.

10.1 Resonators

Resonators are used in a variety of filters and in oscillator circuits for frequency control. Frequency control refers to both the frequency of oscillation and the purity of the output sinusoid. A resonator which has the ability to precisely control the frequency of oscillation has a high Q factor. Note that in general Q is proportional to the inverse of the tuning-range and it is also proportional to resonator size (volume). There are a few types of resonators commonly used at microwave frequencies:

- Lumped element resonators are high-Q capacitors and inductors with associated parasitics. These elements have relatively low values of Q (find some spec sheets to see how large they can be). Also, devices with a capacitance value controlled by an applied DC voltage are available for a tunable lumped element resonator. Oscillators using such a component (e.g. a varactor diode) for frequency control are called VCOs.
- Microstrip resonators can be open or shorted sections of line that provide the right impedance for instability, rectangular $\lambda/2$ resonating line sections, circular disks, circular rings, triangular microstrip etc.
- Metallic cavity resonators are often used for high-Q and convenient mechanical tunability. Waveguide cavity resonators are usually $\lambda/4$ shorted stubs. The output can be coupled out either with a short loop (magnetic coupling) or a short monopole (electric coupling). Often there is a mechanical tuning screw near the open circuit end of the cavity. The lowest order rectangular cavity resonator is a TE_{101} mode, where the width and the length of the cavity are $\lambda_g/2$ at the resonant frequency.
- Dielectric resonators look like an aspirin pill (often called a dielectric puck). These are low cost resonators used externally in microstrip oscillators. They are made out of barium titanate compounds, with relative dielectric constants between 30 and 90. The increased dielectric constant gives high energy concentration, but also higher losses. The mode used in commercially available resonators is a hybrid mode referred to as $TE_{01\delta}$ mode, where the (l,m,n) subscripts usually used in waveguide resonators are modified to (l,m,δ), where the δ indicates that the rod is a bit larger than half of a period of the field variation. The Q factors of these

resonators are often specified at several thousand. The term DRO is used often and refers to a dielectric resonator oscillator.

- YIG resonators are high-Q ferrite spheres made of yttrium iron garnet, $Y_2Fe_2(FeO_4)$. They can be tuned over a wide range by varying a DC magnetic field, and making use of a magnetic resonance, which ranges between 500 MHz and 50 GHz depending on the material and field used. YIG resonators typically have unloaded Q factors of 1000 or greater. Their high tunability and Q-factor as well as small size (excluding the magnetic field control) make them a special case of the tunability/Q/size laws mentioned earlier. The drawback of a YIG resonator is the slower tuning speed. A varactor based resonator is a faster solution if tuning speed is important.
- Whispering-gallery-mode resonators have been used more recently in high precision oscillators. The term "whispering gallery" refers to the field concentration along the dielectric air interface (think of light in a ring of fiber-optic cable). Sapphire is the main dielectric material used for such a resonator because it has the lowest microwave losses for any known solid. In fact Q factors are now quoted at 3×10^5 at room temperature, 3×10^7 at 77 K and 1×10^10 around 10 K.
- It is also important to acknowledge the use of surface acoustic wave (SAW) resonators at sub-microwave frequencies. These types of resonators make use of the mechanical, or acoustic, resonance of a material to achieve high-Q piezoelectric response. Quartz crystals, used from the kHz range up to about 100 MHz, are the workhorse of low-frequency signal generators. There has also been a lot of work recently in mechanical resonators that can reach GHz frequencies, but it is a field in development and not many products exist at this point.
- Optical fiber resonators are also used for high quality factors (very long fibers), but require electrooptic transducers to be used in microwave circuits.

Fig. 53 shows how most of the above resonators compare in terms of realizable Q-factors and rough size (note that this graph is over 20 years old, but I could not find a new one, if you find one, please let me know).

10.2 Oscillators as Multiport Networks

Microwave oscillators can be one-ports, two-ports or three-ports, depending on what active device is used and how it is connected to the rest of the circuit. A one-port oscillator is schematically shown in Fig. 54a, where s is the reflection coefficient of the device and s_L is the reflection coefficient of the load. The oscillation condition is:

$$s_L \cdot s = 1,$$

or in terms of impedances and applying Kirchoff's voltage law,

$$Z_L + Z = 0 \quad \Rightarrow \quad R_L + R = X_L + X = 0.$$

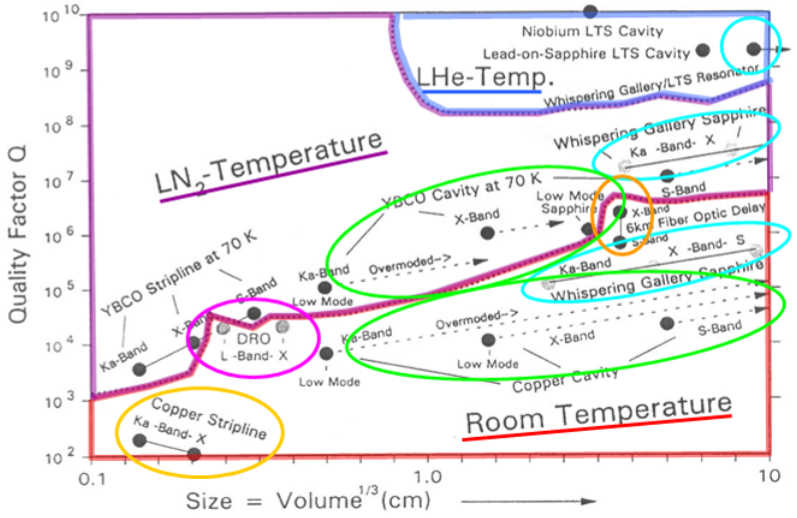


Fig. 53 Comparison of microwave resonators from J. Dick, *IEEE FCS*, 1992.

Since the load is a passive impedance, $R_L > 0$ indicates that R is negative. This makes sense, since an oscillator gives RF power, so that the dissipated power in the oscillator impedance is negative, $P_{osc} = R \cdot I = -|R| \cdot I$. Since any generator can be described as a negative resistor, devices used in oscillators are often called negative resistance devices, although they might or might not have a true negative resistance behavior. Transistors alone do not have true negative resistance, but devices such as Gunn diodes do.

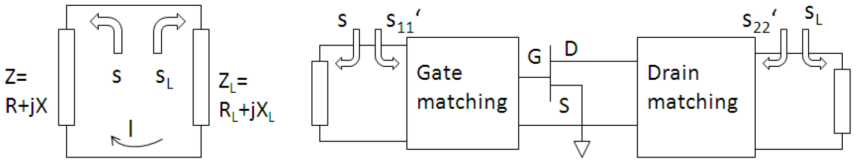


Fig. 54 (a) A one-port and (b) a two-port microwave oscillator block diagram.

The same analysis as above is true for a two-port oscillator. Let us start from Fig. 54b, showing a general amplifier circuit. If there is an RF output with no generator present at port 1, this implies that

$$s = \frac{1}{s'_{11}},$$

and since $|s|$ is less than unity, this implies that $|s'_{11}|$ has to be greater than unity. Consider that in the two-port oscillator circuit, gate matching is a lossless resonator circuit, and drain matching is a lossless matching circuit that enables all of the external RF power to be delivered to the load. Let us assume that the oscillation condition is satisfied at port 1:

$$s = \frac{1}{s'_{11}},$$

where s is the reflection coefficient of the resonator load. We can also write:

$$s'_{11} = s_{11} + \frac{s_{12}\, s_{21}\, s_L}{1 - s_{22}\, s_L} = \frac{s_{11} - \Delta \cdot s_L}{1 - s_{22}\, s_L},$$

so that:

$$\frac{1}{s'_{11}} = \frac{1 - s_{22}\, s_L}{s_{11} - \Delta \cdot s_L},$$

where Δ is the determinant of the transistor scattering matrix. Further, we get

$$s_{11}\, s - \Delta \cdot s_L\, s = 1 - s_{22}\, s_L,$$
$$s_L = \frac{1 - s_{11}\, s}{s_{22} - \Delta \cdot s}$$

At Port 2, we have

$$s'_{22} = s_{22} + \frac{s_{12}\, s_{21}\, s}{1 - s_{11}\, s} = \frac{s_{22} - \Delta \cdot s}{1 - s_{11}\, s}.$$

so that:

$$\frac{1}{s'_{22}} = \frac{1 - s_{11}\, s}{s_{22} - \Delta \cdot s}.$$

Comparing the expressions for $1/s'_{22}$ and s_L, we obtain $1/s'_{22} = s_L$,which tells us that the output port 2 satisfies the oscillation condition as well. A load can be placed in both ports. This result can be generalized to any number of ports, showing that the oscillator is simultaneously oscillating at each of the ports.

10.3 Microwave Oscillator Analysis in CAD Tools

There are several methods that can be used to determine the oscillation frequency in an oscillator:

- examining the open-loop gain of the oscillator circuit;
- examining the closed-loop gain of the oscillator circuit;
- computing the circular function of the oscillator circuit;
- examining the input reflection coefficient at the input/output port of the oscillator.

10.3.1 Open-loop gain analysis

Consider the block diagram of an oscillator consisting of a two-port active device (with $\mathbf{S}^a$) and a two port embedding (passive) network (with a scattering matrix S), Fig. 55. The oscillation condition for a multiport network can be easily generalized from our previous discussion to be:

$$\det(\mathbf{S}^a\,\mathbf{S} - \mathbf{I}) = \mathbf{0}$$

where $\mathbf{I}$ is the identity matrix. It is easy to see how for a one-port oscillator this reduces to our previous condition. For the case of a two-port oscillator, the oscillation condition reduces to:

$$s^a_{11}\,s_{11} + s^a_{12}\,s_{21} + s^a_{21}\,s_{12} + s^a_{22}\,s_{22} - |\mathbf{S}^a|\,|\mathbf{S}| = 1,$$

where

$$|\mathbf{S}^a| = s^a_{11}\,s^a_{22} - s^a_{12}\,s^a_{21} \quad \text{and} \quad |\mathbf{S}| = s_{11}\,s_{22} - s_{12}\,s_{21}.$$

The open-loop gain method of analyzing an oscillator circuit requires that the loop be broken at some point in the circuit, and the loop gain computed over the frequency range of interest. The circuit will oscillate at the frequency where the loop gain has a magnitude greater than unity and a phase of 0 degrees. This condition really means that in order for an oscillation to build up, the total loop gain must be greater than one and the round-trip phase a multiple of 2π. In a circuit simulator, the open loop gain is determined by inserting a probe with the following S-parameters:

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

The idea is that the probe is perfectly matched, lossless and operates as a part of a circulator. This kind of device cannot be implemented in reality. The probe injects a known signal a'_1 which travels around the loop in the clockwise direction, returning to the probe as b'_1. The reflection coefficient seen at port 1' of the probe is the open-loop gain, which can be found to be

$$G_0 = \frac{s^a_{21}\,s_{12}}{1 - s^a_{22}\,s_{22}}$$

There are a few problems with the open-loop gain analysis. First, setting the open-loop gain to be $G_0 = 1\angle 0^\circ$, gives

$$s^a_{21}\,s_{12} + s^a_{22}\,s_{22} = 1$$

which is not the oscillation condition we found above. Another problem of this method is that the frequency of oscillation depends on where the probe is placed in the circuit, what impedance it is normalized to, how it is oriented in the circuit,

etc. All of these problems are consequences of the fact that the probe breaks and therefore changes the loop.

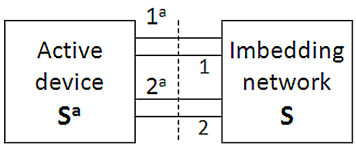


Fig. 55 Generalized 2-port oscillator circuit.

10.3.2 Closed-loop gain analysis

The closed-loop gain approach uses a probe with a scattering matrix equal to:

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

This probe is not physical (since power is not conserved), but it is not invasive, since it maintains closure of the loop. When the closed-loop gain is calculated using this probe, the denominator is exactly the expression from the oscillation condition, which means that when the oscillation condition is satisfied, the closed-loop gain blows up, . The frequency dependence determines whether the circuit can oscillate or not. Briefly, if , the circuit is passive. If and in frequency travels counter-clockwise around the Smith chart, the circuit is capable of oscillation. If the direction is clockwise, it is an amplifier. Since the function tends to infinity, it is not a precise indicator of the oscillation frequency, but rather an indicator of stability. The closed-loop gain is invariant to probe orientation and position in the circuit.

10.3.3 Circular function analysis

If a standard ideal circulator scattering matrix is used for the probe, one gets a circular function for the loop gain as:

$$C = \frac{s_{11}^a\, s_{11} + s_{21}^a s_{12} - |\mathbf{S}^a||\mathbf{S}|}{1 - s_{12}^a\, s_{21} - s_{22}^a\, s_{22}}$$

using a probe matrix

$$\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Under steady-state conditions, when , the expression for C reduces to the oscillation condition for a two-port. The circular function is non-invasive, the result does not depend on its position and orientation and in addition, when looking at a polar plot of C as a function of frequency, it is easy to determine what the oscillation condition is by looking at the intersection of the function with the real axis.

10.3.4 Input/Output *S*-parameter analysis

In circuit simulators such as ADS/MWO, oscillators in Harmonic Balance simulations are analyzed by looking at the open loop gain in the feedback loop. The magnitude of the open loop gain needs to be greater than unity, and the phase zero at the oscillation frequency. If you consider an oscillator from a matched external port, the reasoning is a bit different. You will examine this in your homework and we will discuss it afterwards in class.

A possible oscillator design procedure is:

1. Choose a transistor with enough gain at the desired frequency of operation of the oscillator.
2. Select a resonant circuit and design a topology that has at least one port (e.g. output port).
3. Use transistor small-signal S-parameters at the bias point of interest. Since the resonator mostly determines the oscillation frequency, analyze the reflection coefficient from the port. It will be larger than unity in amplitude if the circuit is oscillating or amplifying.
4. Adjust the parameters of your circuit until you obtain a counter-clockwise loop on the Smith chart, with a radius of the chart between 3 and 10. The counter-clockwise loop indicates an oscillation, as opposed to amplification (look at the S_{21} of any amplifier you designed in previous projects, and you will see that it has a clockwise loop with frequency with a magnitude larger than unity).
5. This analysis is easy to show on a series or parallel resonant circuit connected to a negative resistor and a single port (as done in the RF Lab, ECEN 5634). The small-signal analysis will give you the frequency of oscillation within a few % of the actual one, but cannot predict power, harmonics, waveforms, etc.
6. To obtain the remaining parameters, you will need the device nonlinear model and a harmonic balance analysis (or Spice, in time domain). For HB, you will need to decide what type of oscillator analysis to use (e.g. circular function, etc.). It is good to test that the oscillation condition is met at the same frequency at different points in the circuit.

10.4 Oscillator Example

Figure 56a shows a circuit diagram of a Clapp oscillator topology (modified Colpitts) and a simple first-order equivalent circuit, where it is assumed that the values of C_1 and C_2 either dominate or take into account the intrinsic transistor capacitances.

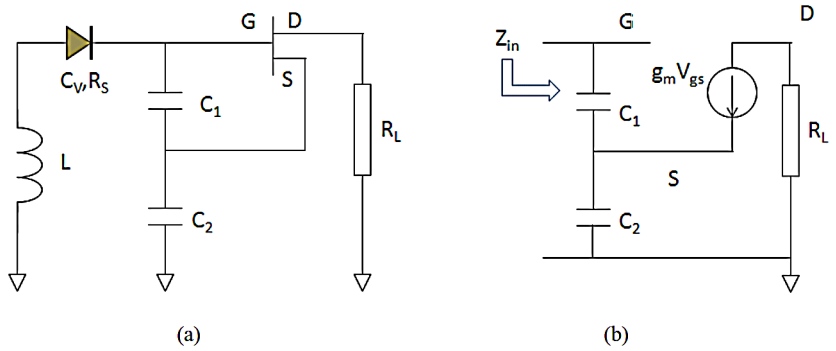


Fig. 56 (a) Clapp transitor oscillator with varactor diode for tuning. (b) Approximate equivalent circuit for transistor with feedback.

By analyzing the circuit in Fig. 56b, we obtain the following approximate expression for the input impedance:

$$Z_{in} \approx \frac{-g_m}{\omega^2 C_1 C_2} - \frac{j(C_1 + C_2)}{\omega C_1 C_2}$$

which shows that the real part of the input impedance is negative, indicating a reflection coefficient larger than unity, while the imaginary part is capacitive, and in fact is the series combination of the two external feedback capacitors. If the series resistance of the external inductor or combination of inductor is R_{LS}, then for the maximal value of $C_1 = C_2 = C_m$, the following condition of oscillation is obtained:

$$\frac{1}{\omega C_m} > \sqrt{\frac{R_{LS}}{g_m}}$$

Therefore, for oscillations to be maintained, the minimal value of the external capacitance is a function of the resistance of the resonator (which could just be an inductor) and the transconductance.

If a varactor diode is added in series to the inductor, the total capacitance is reduced, and the value of the inductor will change for a given resonant frequency. As an example, assume that the large-signal $g_m = 20\,\text{mS}$ and that you chose a varactor that has a capacitance ratio of 4:1 and a parasitic inductance of 0.45 nH. The inductor is implemented with a microstrip transmission line with $Q = 200$ and the capacitances are $C_1 = 2\,\text{pF}$ and $C_2 = 0.5\,\text{pF}$. The dynamic negative resistance is

$$r = \frac{-g_m}{\omega^2 C_1 C_2} = -9\,\Omega$$

This means that the loss in the resonator needs to be smaller than 9 Ω, and this is the combined resistance of the varactor diode and the inductor. A typical medium value for the tuning capacitance is 1 pF. For an assumed oscillation frequency of e.g. 7.5 GHz, this gives an inductance value of about 1.57 nH. From this and the Q factor, we find that the resistance of the inductor is about a quarter of an ohm. Now we can find the tuning bandwidth assuming e.g. 0.4 - 1.6 pF capacitance values of the diode in reverse bias.

Note that it is also possible to use a high-Q resonator, such as a dielectric resonator coupled to the microstrip line, instead of the broadband varactor-tuned version. In that case, the phase noise would be optimized and the oscillator would not be tunable. A dielectric resonator can be modeled as a resonant circuit which is transformer-coupled to the microstrip line. A full-wave simulation is suggested for precise modeling, and varying three parameters can control the Q: distance from end of open or short-circuited microstrip line (L), distance from line that enables coupling (d) and height above substrate (h).

11 Diode Circuits

11.1 Video Detectors

At low RF power levels, a Schottky diode gives a DC output current or voltage that is proportional to the input power level. Since the output is DC, all the phase information is lost. This is called video detection, to distinguish it from coherent heterodyne detection, where a signal is detected with a mixer and a local-oscillator signal, and the phase information in the original RF signal is not lost. The main advantage of video detections is its simplicity. It is not as sensitive as heterodyne detection unless the bandwidth of the RF signal is very large. In terms of sensitivity, it is really only competitive with heterodyne detection for very broadband thermal signals.

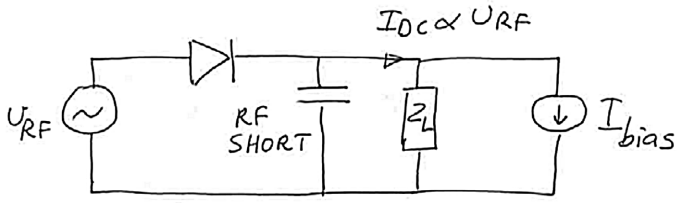


Fig. 57 Schematic of a generic video detector circuit. The output DC current is proportional to the square of the RF voltage.

In a video detector, an RF signal is applied to a diode as shown in Fig. 57. The diode produces a DC output current that is proportional to the signal power, and the current is detected as a voltage across a load resistor. A bias current source can be used to adjust the small-signal resistance of the diode. Consider the IV curve in Fig. 58, with the bias current and bias voltage determining the operating point. When the RF signal is applied, it adds to the bias voltage. Because of the diode nonlinearity, there is a larger change in the current when the voltage is rising than when it is falling. The resulting current waveform has an average value higher than the bias current. The power is detected through this change in bias current. The capacitor in Fig. 57 filters out the RF signal and the harmonics, so that they do not appear across the load.

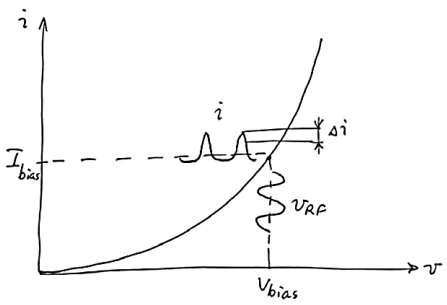


Fig. 58 Diode IV curve showing the video detector operating point and an input RF voltage.

A video detector is characterized by its responsivity, $\mathfrak{R}$, which is the ratio of the output to the input power. We consider two kinds of responsivity: the voltage responsivity $\mathfrak{R}_V$ and current responsivity $\mathfrak{R}_I$, given by

$$\mathfrak{R}_V = \frac{V}{P} \quad \text{and} \quad \mathfrak{R}_I = \frac{I}{P},$$

where I and V are the short-circuited current and the open-circuited voltage and P is the incident RF power. The Norton and Thevenin circuits are shown in Fig. 59. The units for the two responsivities are V/W and A/W. The ratio of the two responsivities is in ohms and represents the video impedance of the detector.

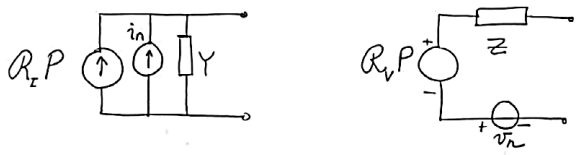


Fig. 59 Thevenin and Norton equivalent circuits for a video detector. The noise sources produced by shot noise in the detector are added in the two models.

In specifying the responsivity, it is important to know what losses are taken into account. For example, we might distinguish between intrinsic detector and system responsivity which might include coupling loss. In addition, it is important to be consistent with the way voltage and power are measured: either both of them with peak-to-peak values (such as with a scope), or both with RMS values (such as with a lock-in amplifier).

Another important characteristic of a video detector is its noise equivalent power. A good responsivity does not necessarily mean a good detector – noise is also important since it determines the signal level that can be detected. For a noisy (realistic) detector, RMS noise sources can be added to the equivalent circuit of the detector. The question then becomes: how much power is this noise equivalent to? The quantity that gives this measure is called the noise-equivalent power and is obtained by referring the noise to the input:

$$\text{NEP} = \frac{i_n}{\Re_I} \quad \text{and} \quad \text{NEP} = \frac{v_n}{\Re_V}$$

If the noise current or voltage comes from the detector itself it will be proportional to the square root of the bandwidth. Recall that the thermal noise voltage from a resistor at temperature T and in a bandwidth Δf is $v_n = \sqrt{kTR}\sqrt{\Delta f}$. This means that the NEP is proportional to the square root of the bandwidth. The bandwidth is the noise bandwidth of the circuit that follows the diode (usually the inverse of the integration time of the amplifier following the detector). Usually a 1 Hz bandwidth is quoted. If the integration time is longer, the noise is lower, but then the measurement is slower.

Typical diode NEPs for microwave GaAs Schottky diodes are around 10^{-12} W/$\sqrt{\text{Hz}}$. The best detectors at microwave frequencies use superconducting detectors, for which the NEP can be as low as 10^{-15} W/$\sqrt{\text{Hz}}$.

For radiometry applications where the signals are low, it is interesting to see what level of temperature change can be detected with a 1-Hz output bandwidth. Assume the input power to the detector is just thermal noise (received, e.g. from an antenna). The received power is proportional to the bandwidth:

$$kT\Delta f = \text{NEP}$$

For an NEP of 10^{-12} W/$\sqrt{\text{Hz}}$, and a 1-Hz RF bandwidth, the smallest detectable temperature is about 10^{11} K (!). This is clearly not a practical detector for radiometry. In order to get a reasonable number, say 1 K, we would need 100 GHz bandwidth, which is not reasonable at microwave frequencies, but is reasonable in the infrared range. However, with a 10^{-15} W/$\sqrt{\text{Hz}}$-NEP cryogenic detector, 100 MHz bandwidth is sufficient to measure 1 K temperature variations (but detector is now not simple).

To analyze the detector, we assume an exponential IV curve:

$$I = I_0 e^{V/V_T}, \quad \text{where} \quad V_T = kT/q = 26\,\text{mV at room temperature.}$$

Then, for $v_{RF} = v\cos\omega_{RF}t$ and assuming that the RF voltage is much smaller than V_T, the diode current can be expanded in a Taylor series. Keeping only the first two terms gives

$$i(t) = I'v\cos(\omega t) + \frac{I''}{2}(v\cos(\omega t))^2 = \frac{I_b}{V_T^2}\frac{1}{2}\frac{v^2}{2} + \frac{I_b}{V_T}v\cos(\omega t) + \frac{I_b}{V_T^2}\frac{1}{2}\frac{1}{2}\cos(2\omega t)$$

The expansion consists of a DC term, a term at the RF frequency and a term at the second harmonic. The RF terms are shorted out by the capacitor, but the DC term gets to the load, and is given by

$$I_{DC} = \frac{I''}{4}v^2$$

The RF power absorbed by the diode is $v^2I'/2$, so that the current responsivity becomes

$$\mathfrak{R}_I = \frac{I}{P} = \frac{I''}{2I'} = \frac{1}{2V_T},$$

the value of which is about 20 A/W. We can write the voltage responsivity as

$$\mathfrak{R}_V = \frac{V}{P} = \frac{\mathfrak{R}_I}{I'} = \frac{1}{2I_b}$$

This means that for a bias current of $100\,\mu$A, the voltage responsivity is a few thousand V/W. Interestingly, the current responsivity does not depend on the bias (although in practice you need some bias to get a reasonable RF impedance so that the following amplifier can work well).

To find the NEP, we assume shot noise in a 1-Hz bandwidth:

$$\text{NEP} = \frac{i_n}{\mathfrak{R}_I} = 2\,V_T\sqrt{2qI_b}$$

Notice that the NEP improves as bias is decreased, as long as the detector can be matched. The NEP can also be improved by decreasing the temperature, since goes down with T. Very good Schottky diode detectors have been made that work at liquid helium temperatures.

In a real diode, there are additional parasitic elements. The most important are the parasitic series resistance R_S and the depletion capacitance C_p. The series resistance is a problem for loss, while in principle the capacitance can be tuned out with an inductor. Both elements depend on bias, and are usually quoted for zero bias. Usually a cutoff frequency determined by these two parasitics is quoted:

$$f_c = \frac{1}{2\pi R_S C_p}$$

although this is just a way to compare diodes, since the diodes typically only work well when the frequency is smaller than about a tenth of the cutoff. Typical numbers

for good GaAs Schottky diodes are $R_S = 20\,\Omega$, $C_p = 20\,\text{fF}$ and $f_c = 800\,GHz$, Fig. 60.

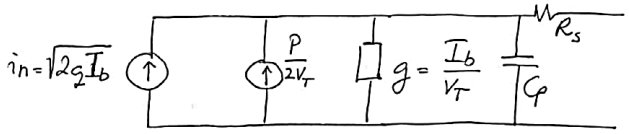


Fig. 60 Equivalent circuit for a Schottky diode detector.

11.2 Resistive Mixers

Mixers are frequency-conversion elements that retain the amplitude and phase of the RF waveform (unlike in a detector, where the phase information is lost). Frequency conversion in a superheterodyne transmitter-receiver system enables inexpensive transmission of a large number of channels on a single, relatively narrowband microwave carrier signal. One negative aspect of introducing a mixer is the added noise.

We start with the simplest mixer analysis: a resistive mixer with a single diode. The inputs to the mixer are an input modulated microwave (RF) signal and a local oscillator (LO) signal, which is generally in the same frequency range as the RF, but has a larger amplitude. The output of the mixer is a signal at the difference frequency between the RF and LO that carries all the information about the input RF modulated signal. The mixer contains a nonlinear device, such as a diode, dual-gate transistor, or (for millimeter-wave frequencies) superconducting resonant tunneling device. There are a number of mixer topologies, depending on the application.

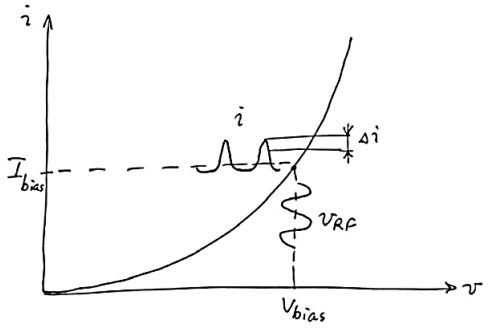


Fig. 61 Diode IV curve showing the mixer operating point and an input RF voltage at a bias point (V_{bias}, I_{bias}.

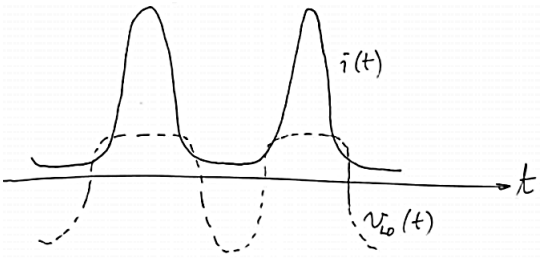


Fig. 62 Current and voltage waveforms in a resistive Schottky diode mixer with only the LO voltage, v_{LO} incident on the diode.

Consider a single Schottky diode as a mixer, and assume that the diode capacitance nonlinearity can be neglected (we will later see that this is a very crude assumption). The nonlinear IV curve $i(t) = f(v(t))$ for the diode is exponential, as shown in Fig. 61. First consider just one signal, called the local oscillator (LO), incident on the diode. The LO is usually a fairly large signal, with about a 1-V amplitude and a power in the few milliwatt range. When a large signal sinusoidal voltage is applied to the IV curve around a bias point in the nonlinear region, the resulting current is a non-sinusoidal waveform. Because of the nonlinearity of the IV curve, small changes in voltage will produce large changes in current when the voltage is in the positive half-cycle, resulting in current spikes, Fig. 62. During the current spikes, the voltage waveform rounds off, because the diode is loading the source. The voltage and current waveforms determine a diode ac conductance which is varying in time with the same period as the LO signal. This time-varying large-signal (or, pumped) conductance can be expressed as

$$g(t) = \frac{\partial I(t)}{\partial V} = f'(V(t)) \tag{116}$$

For an ideal exponential IV curve, we obtain

$$g(t) = \frac{I(t)}{V_T} \propto I \tag{117}$$

Keep in mind that this is just a first-order approximation. One should take the realistic IV curve along with the voltage-dependent capacitance. This complicates the equations, i.e. the shape of the conductance waveform, but does not change the fact that the conductance will be a periodic function with the same period as the LO (pump) voltage. This means that $g(t)$ can be expressed as a Fourier series as follows:

$$g(t) = \sum_{n=-\infty}^{\infty} G_n e^{jn\omega_{LO}t}, \quad \text{where} \quad G_n = \frac{1}{T}\int_{-T/2}^{T/2} g(t)e^{-jn\omega_{LO}t}\mathrm{d}t \tag{118}$$

For a resistive mixer, the G_N's are real, so $G_{-n} = G_n$. Since $g(t) > 0$ (always positive), G_0 is the largest Fourier coefficient. This is easily seen from the following:

$$G_n = \frac{1}{T}\int_{-T/2}^{T/2} g(t)\cos(n\omega_{LO}t)\mathrm{d}t \;\le\; \frac{1}{T}\int_{-T/2}^{T/2} |g(t)|\cdot|\cos(n\omega_{LO}t)|\mathrm{d}t$$
$$\le\; \frac{1}{T}\int_{-T/2}^{T/2} |g(t)|\mathrm{d}t = G_0$$

In the limit, $g(t)$ is a delta function, and all G_n are equal.

The point of the above discussion was to show that an LO (pump) signal has the effect of turning the diode into a conductance which is a periodically varying function with a period equal to the period of the LO signal. When a small-signal (RF) voltage is now in addition incident on the diode, the current through the diode will be the product of the incident voltage and the conductance, $i(t) = g(t)\cdot v_{RF}(t)$. If the RF signal is written as

$$v_{RF}(t) = v\cos(\omega_{RF}t) = \frac{v}{2}e^{j\omega_{RF}t} + \frac{v}{2}e^{-j\omega_{RF}t}$$

where v is real, and the RF frequency is larger than the LO frequency. The current can be written in the following form:

$$i(t) = \frac{v}{2}\sum_{n=-\infty}^{\infty} G_n e^{j\omega_{RF}t+jn\omega_{LO}t} + \frac{v}{2}\sum_{n=-\infty}^{\infty} G_n e^{-j\omega_{RF}t+jn\omega_{LO}t}$$

By observing this equation, the frequencies of the AC current terms are found at

$$\omega_{RF}-\omega_{LO},\;\; \omega_{RF},\;\; \omega_{RF}+\omega_{LO},\;\; \omega_{RF}+2\omega_{LO},\;\; 2\omega_{LO},\;\; 3\omega_{LO}\ldots$$

The first term is the intermediate frequency (IF), which is the desired mixer output. The second term can be used to find the power absorbed by the mixer at the incident voltage frequency, which in turn determines the conversion loss of the mixer. Since the LO is a large signal, all LO harmonics will be present. If in addition the RF signal is large, all harmonics of the RF will be present, along with all mixing terms. The IF term is given by

$$i_{IF}(t) = v\cdot G_1\cos(\omega_{IF}t)$$

and the coefficient G_1 can be viewed as a kind of transconductance of the mixer. The RF (second) terms can be written as

$$i_{RF}(t) = v\cdot G_0\cos(\omega_{RF}t)$$

and the coefficient G_0 represents the effective RF impedance of the device. In purely resistive mixers, this impedance is not a function of frequency, so G_0 is also the output impedance at the IF frequency.

There are two important figures of merit for a mixer: conversion gain and noise figure. The conversion gain for a mixer is defined as

$$G = \frac{P_{IF}}{P_{RF}}$$

and is usually a loss (smaller than unity), with the exception of transistor mixers or negative resistance diode mixers, where it can be a true gain. Usually, an impedance match that optimizes conversion loss does not optimize noise figure. This is derived nicely in the papers by Kerr (posted on the web page).

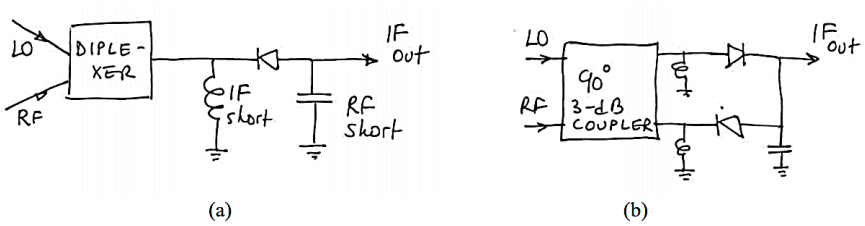


Fig. 63 A single-ended mixer (a) and a balanced mixer (b).

The simplest mixer is a single-ended mixer, Fig. 63a, which consists of a single diode and some kind of a combiner circuit for the LO and RF signals. This mixer has the advantage of being simple, and a disadvantage of LO noise transferring to the IF, and no LO-RF isolation. The LO noise can swamp the small RF signal since the RF frequency is close to the LO frequency in order to make the IF a low-frequency signal easy and inexpensive to process. There are several approaches for reducing the LO noise in the IF: use an LO with very low noise or filter the LO well (expensive), increase the IF (complicates follow-on electronics), or use and additional diode and additional circuits of a balanced mixer, Fig. 63b. The balanced mixer uses a 90° 3-dB coupler which introduces phase shift between the RF and LO signals as they are incident on the two antiparallel diodes. The resulting IF voltage can be written in the form

$$v_{IF} \propto \mathbf{e}^{j\omega_{RF}t}\,\mathbf{e}^{(-jn\omega_{LO}t - j\pi/2)} - \mathbf{e}^{(j\omega_{RF}t - j\pi/2)}\,\mathbf{e}^{-jn\omega_{LO}t}$$

The minus sign is there because the diodes are antiparallel. The two terms add, because the coupler introduces an increase in phase in the first term and an equal decrease in phase in the second term. Consider next a noise voltage in the LO signal. If only the noise that pollutes the RF frequency is considered, the noise voltage can be represented as a narrowband signal $v_n \propto ej\omega_n t$ (an exponential at the RF frequency). The noise voltage passes through the coupler and contributes to the voltage across the two diodes as follows:

$$v_n \propto \mathbf{e}^{j\omega_n t}\,\mathbf{e}^{-jn\omega_{LO}t} - \mathbf{e}^{(j\omega_n t - j\pi/2)}\,\mathbf{e}^{(-jn\omega_{LO}t - j\pi/2)}$$

In this case, the two terms cancel, showing how the balanced mixer reduces the effect of the LO noise. The analysis assumes that the two diodes are identical and the coupler is perfect.

The above resistive mixer analysis did not take into account any nonlinear capacitance or higher-order mixing terms in the diode IV Taylor expansion. The paper by Kerr addresses these issues using a theory which is an extension of the one presented in this lecture, as well as an approach to a more generalized harmonic balance analysis.

11.3 Microwave Rectifiers

There are several applications in which the RF power of a wave is rectified in order to provide DC power for an electronic application. The applications fall into the following categories:

- power beaming, where very high powers are directionally transmitted from one antenna array to another. The receiving array contains rectifying elements (usually diodes) in each element. The polarization is well-known (usually linear) and the bandwidth is essentially zero (single-tone);
- low to medium power distributed wireless sensor powering. In this case, usually the position of the sensors is not exactly known, so the transmitting antenna is not very directional. It is usually operated in a mutipath environment, so polarization and power varies statistically. Therefore, it is useful to rectify any two polarizations independently and add the DC powers.
- Power recycling (scavenging). This is an emerging application and the hardest one since neither polarization, frequency or power levels are known.

In a rectifier, the input RF signal needs to be matched to the rectifying element (diode usually), and the RF impedance optimized at the fundamental and harmonics in order to give the largest power across a DC load. There is an optimal DC load for each incident power level. Since rectification is a nonlinear process, harmonics of the input RF signal are generated and similar techniques as waveshaping in power amplifier design can be used for rectifiers as well in order to increase efficiency. For higher power levels, rectification efficiencies above 80% have been demonstrated at lower microwave frequencies, and on the order of 40% for sub-mW power levels.

MONOLITHIC MICROWAVE INTEGRATED CIRCUITS (MMICS)

1. INTRODUCTION

The world's first MMIC was published in 1975 by Ray Pengelly (later at Cree) and James Turner ("Monolithic Broadband GaAs F.E.T. Amplifiers") in Plessey in the UK. The MMIC, shown in Fig.1, was a single-stage amplifier with 5dB of gain at X-band using 1 micron optically-written gates. The matching circuits were lumped-element designs, but there was no DC blocking on the input/output. Backside processing was not yet possible, so the FET's source was grounded externally.

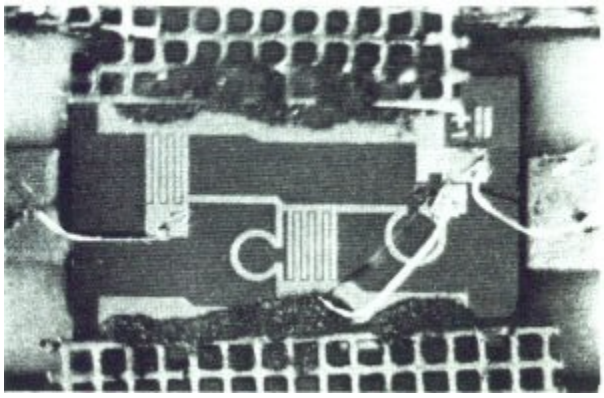


Figure 1. Photo of first MMIC amplifier published in 1975 by Pengelly and Turner.

A MMIC typically refers to a circuit monolithically fabricated in a compound semiconductor, in which the crystal lattice uses two or more types of atoms, such as gallium arsenide (GaAs), gallium nitride (GaN) and silicon germanium (SiGe). Typically there are two types of active devices used in MMICs: vertical devices such as HBTs and PIN diodes, and horizontal devices, such as all types of FETs (MEFETs, pHEMTs, etc.). In a vertical device, each of the semiconductor regions are grown in layers using some type of epitaxy, and the active area is referred to as a "junction". In horizontal (lateral) devices, the active area is referred to as a "channel". These terms are used to define characteristic properties of a MMIC fabrication process. For example, a 0.1-μm FET implies a channel length of 0.1μm.

There are a number of MMIC suppliers currently, most of them in GaAs and some in GaN and InP. The semi-insulating properties of GaAs substrates, and the 12.9 dielectric constant make it an excellent medium for microstrip or CPW design. The loss of GaAs is very low (it can be made to have higher resistivity than high-resistivity silicon). It operates reliably up to 150°C channel temperature and is considered "radiation-hard", which is important for space applications. The main properties of GaAs as a microwave substrate are given below.

Formula or Composition	GaAs
Relative Dielectric Constant	12.88
Dissipation Factor (loss tangent)	0.0004
Temperature Coefficient of relative permittivity	6.8 ppm/°C
Bulk Resistivity	>10E5 Ohm-cm
Mass Density	5.32 gr/cc
Specific Heat	0.33 J/g/°C
Thermal Conductivity (k)	55 W/m°C
Temperature Coefficient of Expansion (TCE)	5.7 ppm/°C
Melting Point	1238°C/2260 °F

2. Brief Overview of MMIC Fabrication Technology

MMIC processing will vary for each foundry and device technology, but many of the steps are common to most processes. In this section, basic fabrication steps are outlined, in order for us to understand the constraints that fabrication imposes on MMIC circuit design.

The overall steps required for semiconductor processing are:

A. Substrate material growth;
B. Wafer manufacturing;
C. Fabrication of active semiconductor surface layers; and
D. Front-side and back-side processing using photolithography.

Photolithography is one of the enablers of IC fabrication. Its goal is to define the patterns for the metal and dielectric components of the IC, and in MMICs it is used for defining transmission lines, lumped elements (inductors, resistors, capacitors) and active devices (transistors, diodes). A mask is first required with the layout in which the metal and dielectric features are created as polygons. UV light illuminates the substrate with a layer of photoresist deposited, and the parts of the photoresist that have been exposed will change chemically, allowing selective removal of photoresist, resulting in a pattern. The mask is typically made of chromium on glass (or iron-oxide on glass). The chromium does not pass UV light used in photoresist exposure. The mask pattern will depend on what type of photoresist is used; positive or negative.

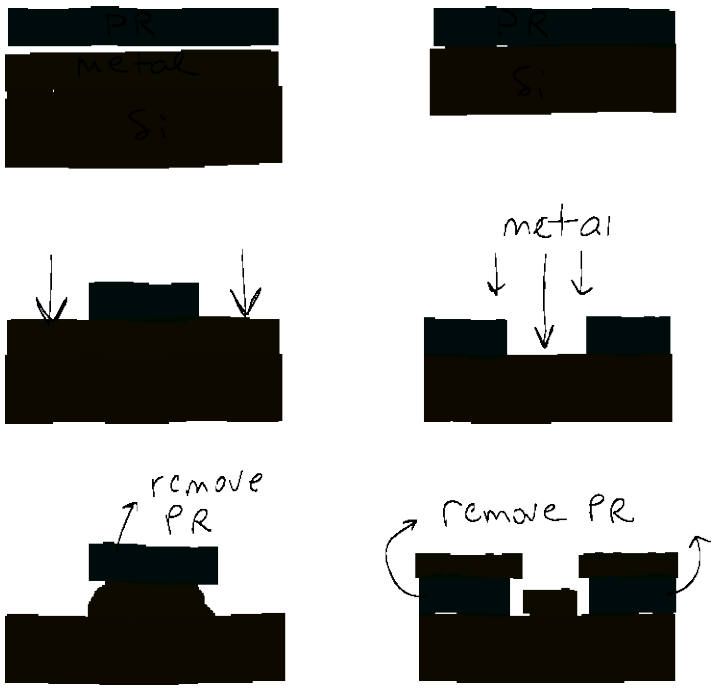
Photoresist is a liquid and is deposited in a thin relatively uniform layer on the substrate by spinning. The thickness of the photoresist, on the order of microns, depends on the density of the photoresist (there are many types) and the spinning speed. The major manufacturers of photoresist and related chemicals (developers, solvents) over the years have been Shipley (bought by Rohm and Haas, now a subsidiary of Dow), DuPont, Fujifilm Electronic Materials, Sumitomo, AZ Electronic Materials and JSR Micro. Negative photoresist is easier to process with a limit of about 2μm in feature size. Positive photoresist is more expensive and has some processing drawbacks, but can define features as small as 0.5μm. In MMICs, gates are smaller

than this feature size at higher frequencies, in which case that part of the mask is written using electron-beam lithography which is more expensive and requires special photoresist.

The metal/dielectric patterns can be formed in two ways:

- Etching, where the metal is etched off where not needed; and
- Lift-off, where the metal not needed is deposited on top of photoresist and taken off when the photoresist is stripped off.

Both methods are sketched in Figure 2.



(a) (b)

Figure 2. (a) Etching process steps for defining metal pattern, and (b) lift-off process steps for defining a metal pattern.

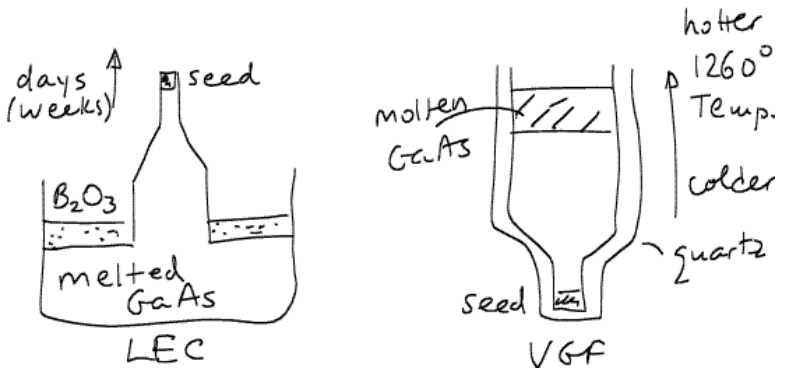
A. Substrate Material Growth

The substrate for MMICs needs to have (1) good insulating properties (high resistivity/low loss) for passive components; and (2) high electron mobility for active devices. Single-crystal GaAs is grown using either a LEC (Liquid Encapsulated Czochralski) process or VGF (Vertical Gradient Freeze) process, which are sketched in Figure 3. The result of this method is a so called "boule" which is a cylindrical single-crystal semiconductor rod. A 4-inch diameter is used by most foundries today.

B. Wafer Manufacturing

The boule next needs to be divided into wafers. Most MMICs end up being 100-μm thick, but the initial wafers need to be thicker for processing, since GaAs is quite fragile (unlike silicon). It is important to identify the crystal axes. The wafers are first diced with a saw, then chemically

etched to make the surfaces flat. This is followed by polishing of the front side to about 2μm flatness. As a result, we now have 4-inch diameter circular wafers of GaAs which is semi-insulating and a very good substrate for transmission lines. The range of resistivities of various semiconductors is given in the table below.



(a) (b)

Figure 3. (a) LEC GaAs growth is done by dipping the seed crystal into molten GaAs where it grows. The seed is slowly pulled out of the molten GaAs and the speed of the pulling determines the diameter of the resulting crystal boule. The molten GaAs is covered with a layer of liquid boron trioxide which prevents volatile arsenic sublimation. (b) VGF growth involves the liquid GaAs in a vertical container with the seed crystal at the bottom, and a temperature gradient is moved up the furnace. This, the single GaAs crystal is freeze-formed vertically upwards. The diameter depends on the size of the mold.

Material	Resistivity range
Si	10^3-10^5 Ωcm
GaAs	10^6-10^9 Ωcm
SiC	10^6-10^{12} Ωcm
InP	10^6-10^9 Ωcm
GaN	10^8 Ωcm

C. Fabrication of Active Semiconductor Surface Layers

The GaAs is undoped at the beginning of this stage, and the active layers can be produced in three different ways:

- ion implantation of dopants, used commonly in Si and for MESFET manufacturing;
- molecular beam epitaxy (MBE), used for HEMTs;
- molecular chemical vapor deposition (MOCVD), used for HBTs.

D. Photolithography

Photolithography is next used to define features on the front (doped) and back side of the wafer:

- Front-side processing includes (1) active device formation; (2) passive component formation; (3) interconnect formation; and (4) passivation.
- Back-side processing includes (5) wafer thinning; (6) through-via hole formation; and (7) back-side metallization.

(1) Active device formation consists of several steps:

- *Device isolation*, required to isolate different active devices (transistors and diodes). This can be done through mesa etching (forming islands of doped material isolated by a sea of semi-insulating GaAs), or by ion implantation which breaks the crystal structure between active devices, making amorphous semiconductors which do not conduct. Mesa etching is usually wet etching along crystal planes, resulting in a different angle along different cuts, as sketched in Figure 4. Mesa resistors are also formed in this step of the process, with about 300Ω/□ surface resistivity in the case of GaAs. This process results in non-planar surfaces after isolation, while the ion implantation results in a flat surface but it is a high-energy process and cannot isolate to any depth.
- *Ohmic contact formation* consists of high doping (n+ or ~ $10^{18}/cm^3$) followed by evaporation and lift-off to deposit alloys (gold, germanium) which form a very low-resistance contact after they are baked to form and stabilize the alloy. This is used to define the source and drain metallization on the first metal layer.
- *Gate (Schottky) contact formation* results in a low-doped contact which forms a diode, and is done by deposition of titanium, platinum and gold layers (platinum is used so that gold does not diffuse into the GaAs through the thin titanium layer). This is also the first interconnect layer, and the total thickness of the 3 metals is on the order of 0.5μm.
- *The gate patterning* is often done with electron-beam lithography with special photoresists. At this point, DC testing for pinch-off voltage and drain current is typically done.

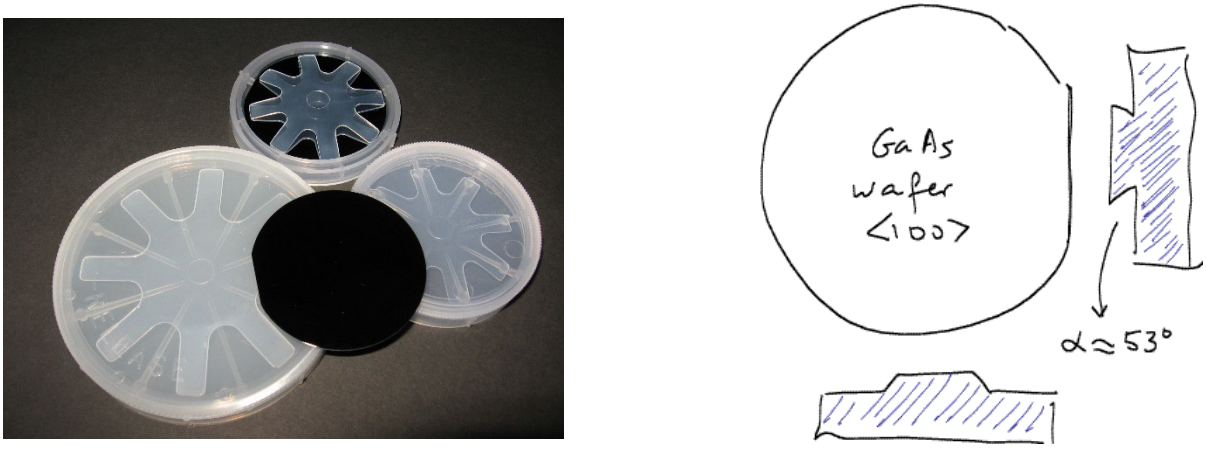


Figure 4. Wafer with crystal axis marked by flat side. Mesa etching along crystal planes leaves asymmetrical mesa shapes. The ones along the horizontal cut are convenient for metallization and result in good continuous contacts, while the perpendicular mesas form shadows for metal deposition. Design rules in this case will dictate that active device gates are oriented in a specific direction allowing for good metallization.

(2) Passive component formation consists of several steps as well:

- *High-permittivity dielectric deposition* and patterning for capacitors, using PECVD (plasma chemical vapor deposition) of silicon nitride (SiN_4) with a permittivity of 6.8. Capacitors made in this way have higher capacitance values. The dielectric is typically about 1000 Angstroms thick and has a breakdown voltage around 50V.
- *Resistive material deposition* and patterning for thin-film resistors is done next, using usually NiCr thin films.
- *Low-permittivity dielectric deposition* and patterning for cross-overs and small capacitors is done next, using polyimides (permittivity of around 2.7), typically 3-5μm thick.

(3) Most of the interconnects are either made together with the gate contact, or with metal *sputtering of a thick layer for thermal shunts and interconnects*. This is metal layer 3.

(4) Following the interconnect layer, *passivation* of the wafer is done with silicon nitride deposition, but lines for scribing or dicing must be left as SiN_4 leaves cracked edges.

(5) Back-side processing starts with *wafer thinning* to usually 100μm thick, but can be 50μm for millimeter-wave devices. Polyimide coating is used for front-side protection during this process. The wafer is placed with wax on glass and then lapped and polished to avoid losses due to surface roughness.

(6) The next step uses *reactive ion etching (REI) to form via*s (holes) through the GaAs from the front to the back side. The diameter of the vias is kept typically below half of the wafer thickness. This step involves more patterning and mechanical protection of the wafer.

(7) Finally, *back-side metallization* with thick gold (for soldering and bonding) defines the ground for the microstrip and fills the vias with metal to form ground connections.

After this, the wafer is mounted on a tacky-film, and then scribed. When the tacky film is heated, it expands, leaving the individual MMIC chips.

Below is an overview of the main types of active devices used in MMICs. This is followed by an overview of passive MMIC devices and packaging techniques.

3. ACTIVE MMIC DEVICE OVERVIEW

1. GaAs FETs

The first 2-inch wafer GaAs MMICs demonstrated in the 1970s were based on MESFETs. GaAs MESFET MMICs will never be cheaper than silicon, due to the starting material cost related to more expensive fabrication. GaAs parts are more fragile than silicon, and the thermal dissipation factor is not as good. GaAs MESFETs are now largely replaced by pHEMTs (pseudomorphic HEMTs), offering higher performance and are not much more expensive to fabricate. The GaAs pHEMT was the second MMIC technology to be perfected, in the 1990s. Breakdown voltages of pHEMT up to 16V make high-power/high efficiency amplifiers possible, and excellent LNAs with noise figures in the tenth of a dB range at X-band are possible. pHEMT stands for pseudomorphic high electron mobility transistor. The semiconductor is not just GaAs, but in addition, e.g. AlGaAs/InGaAs/GaAs, which means the pHEMT is a hetero-structure.

“Pseudomorphic” means that the hetero layers are thin enough not to keep their own crystal lattice structure, but assume the structure (lattice constants especially) of surrounding material (implying a high level of stress between layers). In a two-dimensional cross section of the layer, it can be seen that, while it assumes the lattice constant of the bulk structure in the X direction, it tries to keep its original lattice constant in the vertical direction (so it is really strained). For a GaAs pHEMT, indium is added to improve mobility and form a quantum well.

MHEMT stands for metamorphic high-electron mobility transistor. The channel material is InGaAs. "Metamorphic" implies that the lattice structure of GaAs is buffered using epitaxial layers to gradually transform the lattice constant so it lines up with InGaAs. InGaAs is normally grown on InP, which is expensive and fragile compared to GaAs. “Metamorphic” is changing the lattice constant by bond breaking, so it is very strained. These devices can work up to 100GHz and noise figure and f_{max} can equal that of indium phosphide if indium content is high. Below are listed some advantages and disadvantages of GaAs MESFETs, pHEMTs and MHEMTs.

MESFET Advantages:	MESFET Disadvantages
• Mature technology • Optical gates (usually) means low cost • Great microwave substrate (relative permittivity of 12.9, low loss tangent, high bulk resistivity) • Six inch wafers available • Photonic properties • 16-20 volt breakdown possible • Relatively cheap to produce (but always more than silicon) • Channel temperatures up to 150C possible	• Limited to Ku-band or lower • Noise figure and power performance not as good as GaAs PHEMT • Positive and negative voltage typically needed (VGS and VDS).

pHEMT Advantages:	pHEMT Disadvantages
• Useful through Q-band, especially if thinned to 2 mils and individual source vias are used • Excellent power and efficiency (greater than 60% PAE) • Breakdown 12 volts at best, typical operate at 5-6 volts • Channel temperatures up to 150C possible.	• E-beam gates (increases cost) • Positive and negative voltage typically needed (VGS and VDS)

MHEMT Advantages:	MHEMT Disadvantages
• Extremely low noise figure • Incredibly high fmax (more than 100 GHz) • Extremely low on-resistance, makes great switches, but not as good as PIN diodes. • Channel temperatures up to 150C possible.	• Breakdown voltage much lower than PHEMT • Low operating voltage (1 to 2 volts) • Positive and negative voltage typically needed (VGS and VDS)

2. GaAs and other HBTs

The heterojunction bipolar transistor (HBT) can decrease the cost of GaAs amplifier products because the emitters are formed optically, possible due to the vertical operation. However, for very high frequency, the emitter size must be small, and the InGaAs layer is thick and is a thermal insulator, so these devices tend to run hot. Typical HBT amps are gain stages, used in the UHF to C-band frequency ranges. Some processes allow for both HBTs and pHEMTs on the same wafer.

Advantages:	Disadvantages
• Single power supply polarity • All-optical process	• Heat dissipation can be problem at small emitter size • Typically, reverse isolation is not as high as with PHEMT amplifiers, leading to poor amplifier directivity. • Collector resistors are required to stabilize amplifiers. These cut into your power efficiency.

3. GaAs VPIN diode

PIN diodes make great switching elements, as well as limiters. Vertical PINs (VPINs) are offered on some MMICs, but VPIN diodes and amplifier devices such as FETs on the same wafer are not made. M/A-COM and Qorvo (Texas) make VPIN MMICs.

Advantages:	Disadvantages
• The lowest on-resistance for the least amount of off-capacitance. • Huge power handling.	• Two terminal device means you must create bias tees to bring in DC control signals. • Expect DC current up to 20 mA to create a good RF short circuit.

4. Gallium nitride (GaN)

More expensive in terms of dollars per die, GaN offers a path to much higher power densities and therefore cheaper power. Breakdown voltages of over 150 Volts are possible, and one can purchase 50-V parts at X-band. GaN is still a somewhat immature process, though recently space qualified. Reliability has been a problem that is just being overcome, and thermally induced trapping effects lead to circuit nonlinearities.

Substrates for GaN are either SiC, sapphire, or silicon (MaCOM and Ommic in France use this approach). Native GaN wafers are impractical, so a lot of expensive processing is needed to align the GaN crystal onto mismatched substrates. SiC is an excellent heat sink, and GaN can operate up to greater than 150°C channel temperature. Below 2 GHz, GaN is used in base station applications, competing with LDMOS technology. Higher frequency GaN products are now available in a limited way at millimeter-waves (W-band). Silicon is not as good of a heat sink as silicon carbide (40 versus 350 W/m-K), so lower-cost of GaN on-silicon-may be outweighed by the ability to dissipate higher power (and thereby achieve greater power density) on SiC.

GaN Advantages:	Disadvantages
• Up to 10X the power density of GaAs PHEMT has been demonstrated. • Higher operating voltage, less current. • Excellent efficiency possible. • SiC substrates are great heat spreaders. • Can operate hotter than GaAs, Si or SiGe.	• Expensive • Reliability not established yet • Large heat flux

5. Silicon and SiGe

SiGe is low cost due to processing on eight-inch (200 mm) diameters wafers. However, the devices do not have as high of performance as GaAs, in terms of noise figure and power, although they do work at very high frequencies. The setup charge at IBM to make a mask set is enormous, because 200 mm contact masks are needed (GaAs usually uses a 10X wafer stepper, these glass reticles are relatively cheap). The poor insulating properties of a silicon substrate means it is not a good medium for microstrip, so one can make transmission lines in the backend of line (BEOL) SiO2 and metal layers. The SiO2 dielectric layers are thin, which means high metal losses. Another option is to send your wafers to a third party for post-processing to put a lower dielectric metal system on top of it, such as benzo-cyclo-butene (BCB) and gold. When the upper frequency of SiGe extended (up to 300GHz fT in the IBM/Global Foundries 6th generation), the breakdown voltage is reduced, so power amplifiers are not made in SiGe. However, the phase noise is low, so they make good oscillators.

4. PASSIVE MMIC DEVICE OVERVIEW – LUMPED ELEMENTS

1. Via holes from top metalization to ground

The inductance of a single via to ground can be calculated as in the IEEE paper "Modeling Via Grounds in Microstrip" IEEE Microwave and Guided Wave Letters, Vol. 1, No. 6, June 1991, by Goldfarb and Pucel. This is an empirical solution that was found to fit data where the height was between 100 μm and 631 μm. It is supposed to be valid for heights less that 3% of a wavelength.

$$L_{via} = \frac{\mu_0}{2\pi} \cdot \left[h \cdot \ln\left(\frac{2h + \sqrt{r^2 + (2h)^2}}{r} \right) + \frac{3}{2}\left(r - \sqrt{r^2 - h^2} \right) \right]$$

If the units of all of the lengths are in meters, then the inductance is in Henries. The inductance of a via hole for a standard 4-mil thick (100-μm) substrate is shown in Figure 5, where *D* is the diameter (*r* is the radius in the formula), and *h* is the substrate height. In many cases it is desirable to keep the *D/H* ratio to 0.5 to minimize the catch-pad area, so 20pH is a good number to use as a guidance in circuit design.

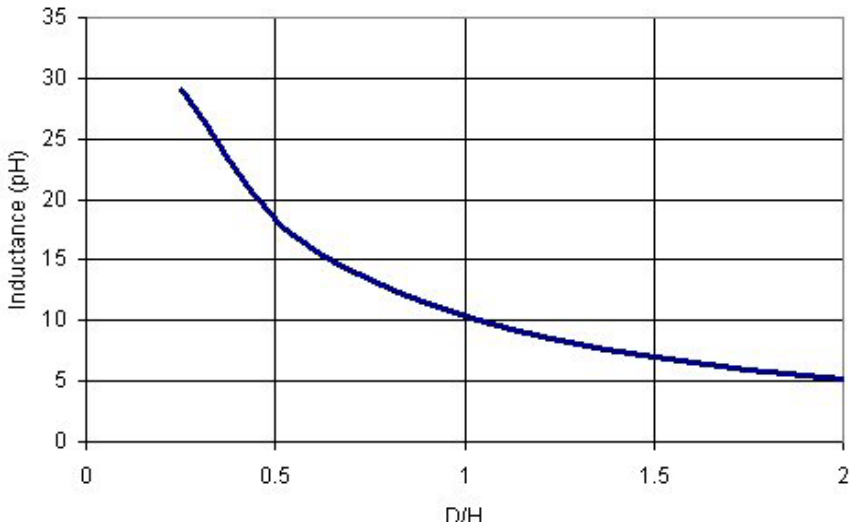


Figure 5. Inductance of a via as a function of ratio of via diameter to substrate height for GaAs. It is advisable to use ratios less than 0.5 so that the pads on the front side connecting to the vias will have reasonable size with low parasitics.

2. Resistors: thin film and mesa types

Thin-film resistors can be implemented by depositing NiCr (nickel chromium) or TaN (tantalum nitride), both with surface resitivities of around 20-50Ω/□. The advantage of NiCr is that its resistivity does not change with temperature. For straight resistors (uniform current density approximation), it is easy to make a unit cell and cascade it to get different values. An example using a 50Ω/□ cell is shown in Figure 6. Note than when a resistor is meandered, this method is not valid and a full-wave simulation is required. Also, if the resistor becomes an appreciable electrical length ($\lambda_g/20$ or so), it needs to be treated as a lossy transmission line section, as opposed to a lumped element. NiCr resistors can handle 0.5mA of current for every micro-meter of width.

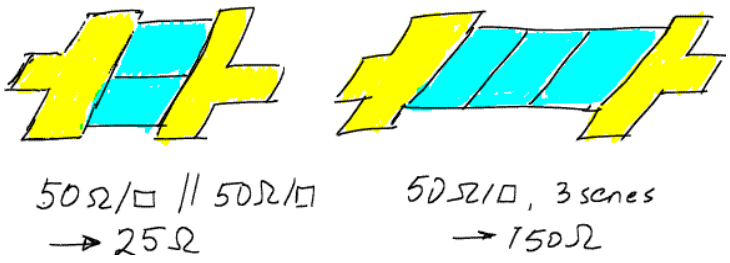


Figure 6. Examples of obtaining different values of resistors using a standard square.

The resistor value does not depend on the size of the resistor, i.e. a 20µm by 20µm resistor has the same value as a 100µm by 100µm resistor if they are both made of a film with the same surface resistivity (why?), but the current and power handling capability will be different. How would you find the power handling capability of a resistor for which you know the size (area) and surface resistivity? Is there any other material property you need to know?

Mesa resistors are used in MMICs to provide high-value resistance, but with low accuracy (tolerance +/-20%), limited power handling and high temperature coefficient of resistance. This is useful for bias lines and is often used in controlling the gate electrode of a switching FET. A mesa is an area on a semiconductor wafer where the semiconductor has not been etched away. In nature a mesa is a flat-topped mountain; on a semiconductor a mesa also rises above the surrounding semi-insulating substrate, but the height is typically less than one micron. It is possible to create a resistor using ion-implant techniques that are purely planar to the substrate, and have the same properties as mesa resistors.

Contacts to mesa resistors are established using source-drain ohmic metal. A passivation layer is always added to protect the resistor (SiO_2 shown in pink in Figure 7) and has to be opened up to allow contact between the thick metal used for transmission lines and the ohmic metal which should be avoided for transmission lines because it is lossy. The total resistor value has three series components which we have denoted RC, RM and RC. The mesa itself (RM) has a value that is calculated from its sheet resistance and number of squares chosen by the designer (length/width). Typical values for sheet resistance are 100-500 Ω/square depending on what process you are using, which is usually much higher than you can get with thin-film metal resistors. The RF sheet resistance is essentially the same as the DC resistance, because the thickness of mesa resistors is typically much smaller than a skin depth.

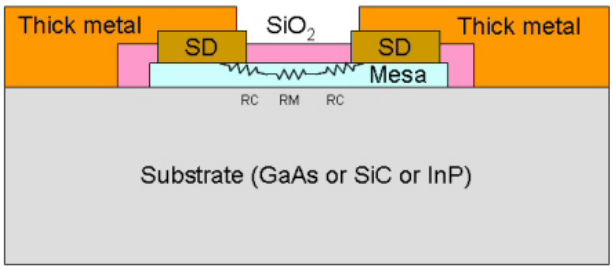


Figure 7. Cross-section of a mesa resistor showing different parts that contribute to total resistance.

The contact resistance "RC" shown in the cross section is usually given in Ω-µm, and you calculate its value by dividing by the width of the contact, not the area. In this case the only dimension you have to consider is the width of the contact region, not the entire contact surface. Typical values for contact resistance are 100 to 300 Ω-µm, so a 10-µm wide contact would have 10-30 Ω of contact resistance. Contact resistance is typically measured or calculated for direct current only, but it is a reasonably good approximation for RF as well. There is a property of mesa resistors that you need to consider if you want to pass any appreciable amount to current through them. These are semiconductor resistors, which means they are not linear. The parameter that is often associated with the nonlinear behavior is the critical field voltage (V_{crit}) expressed in V/µm. However, it is far easier for a circuit designer to think in terms of the saturation current per unit width: $I_{sat} = V_{crit} / R_s$, where R_s is the sheet resistance for the process. The sketch in Figure 8 illustrates this with an example of a sheet resistance of 200Ω/□ and V_{crit}=0.3V/µm. Therefore any resistor of 10µm width will saturate at 1.5mA. The example resistor is 10µm wide by 100µm long, which should be good to about 30V. Mesa resistors have very poor temperature coefficients, at around -0.1%/degree C. Over a 100 degree C range, your resistor will vary 10%!

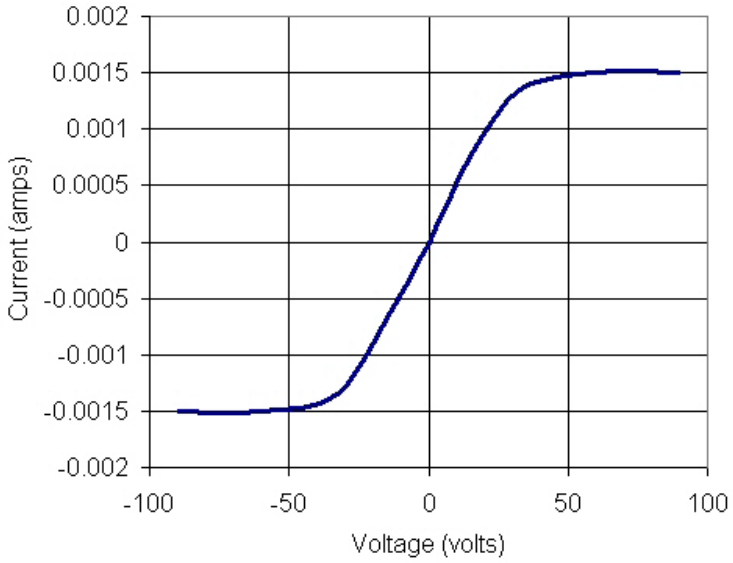


Figure 8. Sketch of the IV curve for a mesa (semiconductor) resistor. The example is calculated for a sheet resistance of 200Ω/□, 10μm wide by 100μm long.

Capacitors

There are two kinds of MMIC capacitors:

- interdigitated and
- metal-insulator-metal (MIM) parallel-plate.

An interdigitated capacitor does not need an extra dielectric layer, but the capacitance values are limited to less than 1pF and the Q factors are low compared to the parallel-plate capacitors, since the fields penetrate into the GaAs. MIM capacitors can vary in size from 20μm by 20μm to 100μm by 100μm and the value can be from 50fF to 200pF. The dielectric used for larger capacitance values is SiN_4, typically 100-120nm thick. For lower values, polyimides (permittivity 4.5) and BCB (permittivity 2.7) with thicknesses of 1-3μm are used. SiN_4 has the lowest loss and highest breakdown voltage (65-85V).

Inductors

A section of transmission line is an inductor, and the value of inductance is larger for a higher impedance line, as illustrated with some typical numbers in Figure 9. Generally, the value of inductance is around 1nH/mm and the DC current handling capability is around 10μm per micro-meter of width. This value can be estimated quasi-statically using image theory, which results in a two-wire strip line. The equivalent radius of a circular wire is $w=4r$, where w is the strip width. The static formula for inductance per unit length becomes

$$L' = \frac{\mu_0}{\pi} \ln \frac{d-r}{r} \approx \frac{\mu_0}{\pi} \ln \frac{d-w/4}{w/4}$$

and gives values within a factor of 2 for different metal thicknesses.

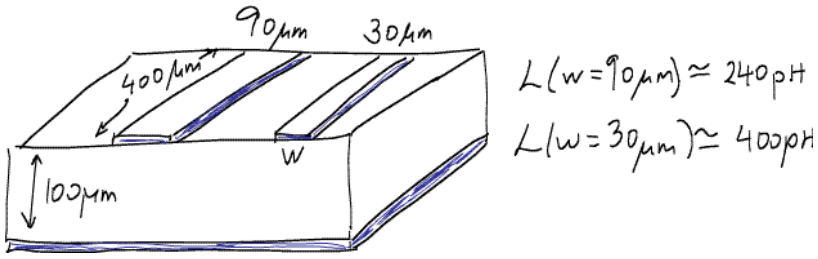


Figure 9. Sketch of inductance of microstrip lines of two different widths on a standard 100-μm thick substrate.

Equations that match experiment better have been published for air bridges and bond-wires, below is a formula used often for a ribbon air bridge, where all dimensions are given in mils, L is the length, W the width and T is the thickness of the metal:

$$L(nH) = 5.08 \cdot 10^{-3} L \cdot \ln \frac{L}{W+T} + 1.19 + 0.222 \frac{W+T}{L}$$

The formula can be found, e.g. in *Advances in Microwaves*, Volume 8, Academic Press 1974, as well as in *Computer-Aided Design of Microwave Circuits*, by K.C. Gupta, Ramesh Garg and Rakesh Chadha, Artech House 1981. Gupta's book mentions the inductance of ribbon strips going down as frequency increases, due to surface reactance. That reactance goes up only as the square root of frequency, which means the surface inductance is going down with the square root of frequency. At some frequency that term becomes insignificant and the other term takes over. The book refers to "Calculation of Inductance of Finite-Length Strips and Its Variations with Frequency", IEEE MTT, Vol. MTT-21, 1973.

A variety of spiral inductors are made on MMICs for increased values of inductance, and these require a minimum of two metallization layers, as seen in the sketch in Figure 10. It is also possible to increase the magnetic flux by stacking two inductors as shown in the figure. Values of inductors that can be implemented at microwave frequencies in a MMIC range from a fraction of a nH to tens of nH. More windings have higher inductance, but also higher capacitance between the traces, resulting in a lower resonant frequency. Three parameters are important for specifying an inductor: value of inductance, resonant frequency and Q factor. The inductance value is valid until approximately half of the resonant frequency, which is where the inductor is a parallel resonant circuit with infinite impedance and after which the inductor behaves as a capacitor. The value of the capacitance can be estimated based on the simple resonant circuit formula. The Q factor is approximately $\omega L/R$, so the value of the series resistance can be estimated from this value. The losses in an inductor come from the resistive losses associated with skin depth in the metal (gold), but these are increased at higher frequencies due to the proximity effect which results in higher current density at the edges of the traces and at the corners. Thus, circular inductors have higher Q than square ones, but the strip width can be varied in square inductors to compensate for this. MMIC design rules give good guidance for inductor design if you use a standard inductor shape.

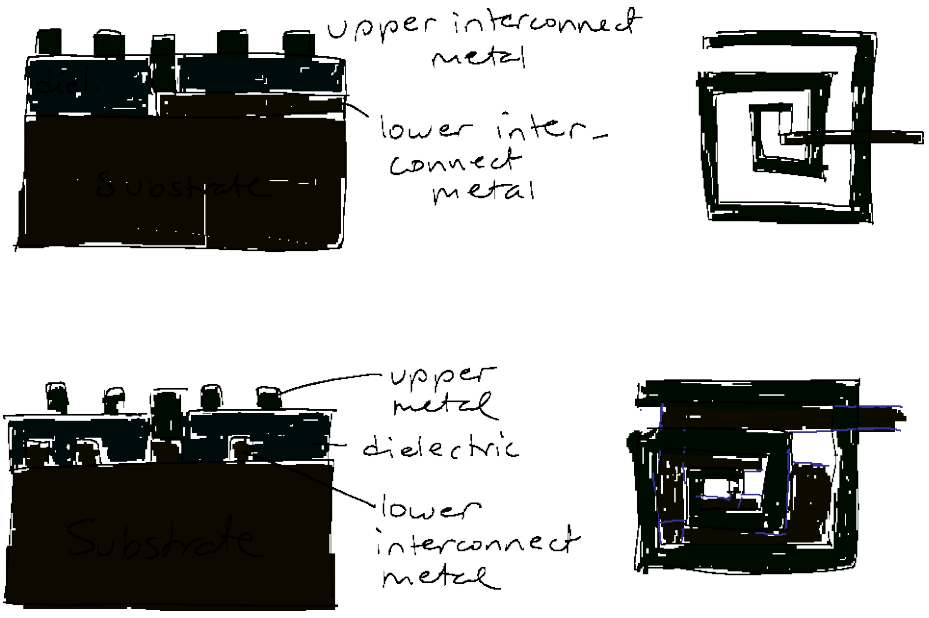


Figure 10. Sketch of spiral inductors using 2 metal layers: single-spiral inductor (top) and dual- spiral (bottom) inductors. The top has less inductance but a higher resonant frequency.

Transmission lines

Semi-insulating GaAs is an excellent substrate for transmission lines. Microstrip is very common and a range of impedances is possible. The cross-section of a MMIC microstrip is not exactly like the textbook type, Figure 11, so there are modifications when calculating the impedance. The effective dielectric constant of the lines is usually determined empirically and matched to a 2.5-D electromagnetic numerical model. Resonators such as the one shown in Figure 12 can be (and are) used to determine the effective dielectric constant. Several such resonators with different characteristic impedances in the ring are fabricated and measured, since the effective dielectric constant is a function of width of line (impedance). The effective dielectric constant ranges from 7.5 to 8.6 for width-to-substrate heigh ratios between 0.05 and 1, respectively.

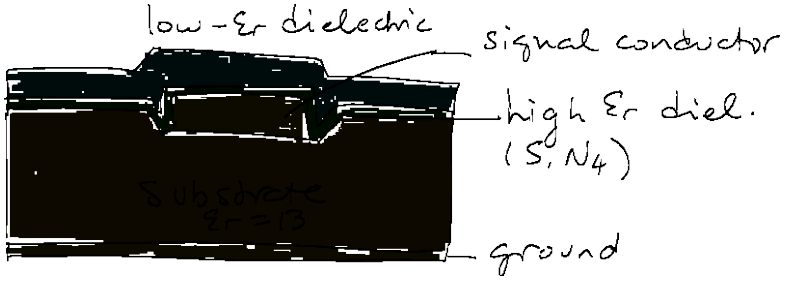


Figure 11. Sketch (not to scale) of a cross-section of a MMIC transmission line.

The range of characteristic impedances is from about 120 to 40Ω, and the impedances are higher for transmission lines which use the top metallization layer for the signal line (why?). Depending on the substrate thickness, the microstrip lines are a few μm to 100μm wide. All passive components that we have discussed so far can be implemented in MMICs, and some of them are easier in MMICs than in hybrid circuits (e.g. Lange couplers). The loss for microstrip on GaAs is in the range of 0.05dB/mm and 0.2dB/mm and depends on frequency and impedance, as well as metal layer the line is implemented in (the thick top layers result in lower loss).

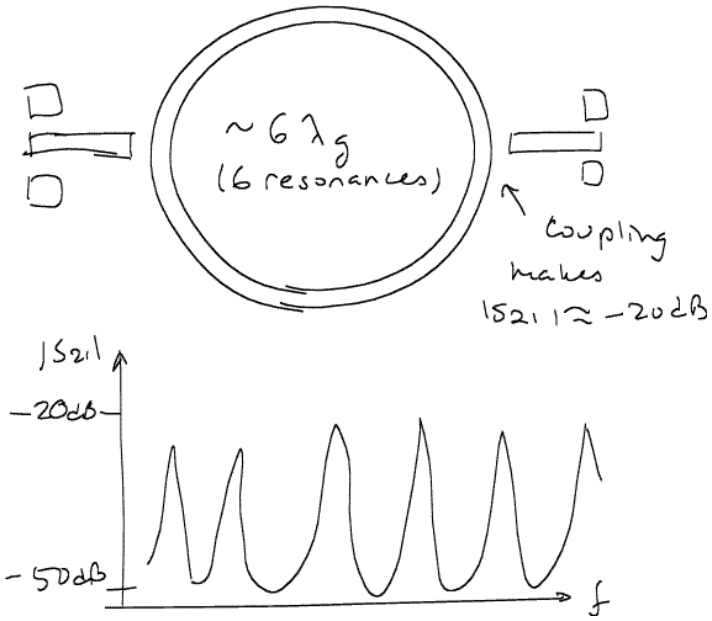


Figure 12. Sketch of circuit used to determine the effective dielectric constant of a MMIC transmission line (top) and typical transmission response (bottom). The effective dielectric constant can be measured from the peaks of $|s_{21}|$.