Contents

Chapter 1 Importance of Impedance Matching

1.1 Difference between RF and Digital Circuit Design

It is well known that the unit of digital data rate is bps (bits per second), Mbps (Megabits per second), or Gbps (Giga bits per second) while the unit of frequency is Hz (Hertz), MHz (Mega Hertz) or GHz (Giga Hertz). They are comparable because they have the same dimension, 1/second. In respect to RF (radio frequency), a digital circuit can be distinguished into two cases: low and high digital data rate.

A digital circuit with a low data rate can be defined as

$$R << f_{RF} \ , \tag{1.1}$$

while a digital circuit with a high data rate can be defined as

$$R \approx f_{RF} \ , \quad \text{and} \quad R > f_{RF} \ , \tag{1.2}$$

where the digital data rate is denoted by R and the frequency of the RF signal is denoted by f_{RF}.

The RF range has not been well-regulated and confirmed worldwide. Approximately, they are between MHz to GHz and are changed from time to time. If the low limit of f_{RF} is assumed to be

$$f_{RF} \approx 10 \text{ MHz} \ , \tag{1.3}$$

then, the condition (1.1) to represent the low data rate cases can be re-defined as

$$R << 10 \text{ Mbps} \ , \tag{1.4}$$

which implies an approximation that a digital waveform with 10 Mbps of data rate can be mainly characterized by its main spectral component of 10 MHz.

And, the condition (1.2) to represent the high data rate cases can be expressed by

$$R \approx 10 \text{ Mbps} \ , \text{ and } \quad R > 10 \text{ Mbps} \ . \tag{1.5}$$

In the RF circuit design, special requirements, such as impedance matching, RF grounding and so on, must be always complied with. In the digital circuit design, on the other hand, the design rules are quite different between two cases of low and high data rate. More detailed discussions would be conducted in the following sub-sections.

1.1.1 Case #1: Digital Circuits at Low Data Rate

In the second half of 20th century, the electronic industry has been progressing faster and faster. Various digital and RF circuits have been developed to promote the computer and other electronic products. In the early stages, the digital data rate in a transceiver was much lower than the frequency of an RF signal as described by the definition (1.1) or expression (1.4). One of examples is the wireless communication system in its early stages. It consists of both RF and digital portion, in which the digital data rate is in the orders of Kbps to Mbps while the frequency of an RF signal is in the orders of 10MHz to GHz. The circuits operating for these two types of signal are quite different either on the appearance or in the design philosophy. Table 1.1 summarizes these differences.

Table 1.1 The differences between RF and digital circuits when the data rate is low.

Item	RF module /RFIC	Digital circuit (Low data rate)
Impedance	**Low (50 Ω typically)**	**High (Infinitive ideally)**
Impedance matching	**Important**	**Don't care (usually)**
Current	**High (*m*A)**	**Low (*μ*A)**
Location in the wireless communication system		
* *Rx*	**Front end (Before de-modulator)**	**Back end (After de-modulator)**
* *Tx*	**Back end (After modulator)**	**Front end (Before modulator)**
Transportation type	**Power (Watt)**	**Status (Voltage)**

First, the impedance is quite different. The input and output impedances are usually pretty low in the RF circuitry. They are typically 50 Ω in most test equipments. On the contrary, the input and output impedances are usually high in the digital circuitry. For example, the input and output impedances of an Op-amp(Operating Amplifier) are mostly higher than 10 kΩ.

Secondly, in RF circuit design, either input or output impedance is required and emphasized to be impedance-matched. That is, the input impedance must be matched to the impedance of the source while the output impedance must be matched to the impedance of the load. Impedance matching is one of the key important criteria in the judgment of the correctness of an RF circuit design. On the contrary, in the digital circuit design, the impedance matching is never to be mentioned or to be taken care of. It seems to be an ambiguous or a strange phenomenon. Is it due to academic prejudice or the different methodology in the engineering design?

Third, in RF circuit blocks the current drains are usually in the order of milli-amperes while in the digital circuit blocks they are usually in the order of micro-amperes. That is, the difference of current drain's magnitude between RF and digital circuit block is approximately 1,000 times.

As a matter of fact, from the comparison above it can be concluded that the key difference between RF and digital circuit design is the necessity of impedance matching. It is well known that one of the impedance matching objectives is to reach the maximum of power transportation. It implies that power transportation is one of main functions for an RF circuit since the impedance matching is emphasized. In an RF circuit block, lower impedance and higher current drain are preferred because they are beneficial to power transportation. On the contrary, a digital circuit with low data rate does not perform power transportation since it does not require impedance matching. What we concern of digital circuit is how it manipulates or transports the status of "0" and "1". In other words, a digital circuit manipulates or transports the status but not the power of the digital signal. As long as the "0" or "1" can be manipulated or transported well, it is desired to reduce the power of the digital signal as much as possible. It implies that lower current drain and higher impedance are preferred in digital circuit blocks with low data rate because they are beneficial to the power saving.

Summarily, when a digital circuit is operated at a low data rate, the digital circuit is working for the **Status** transportation or manipulation while an RF circuit is working for the **power** transportation or operation. Either digital or RF circuit design bears their own special task and therefore has their own special features.

The question raised up is: why is power transportation required for an RF signal whereas status transportation asked for a digital signal? Could this requirement be exchanged in the actual engineering circuit design? The answer can be found from the block locations in a wireless communication system as shown in Table 1.1. The interface block between RF blocks and digital blocks is the modulator in the transmitter and the de-modulator in the receiver. In the transmitter, the digital signal modulates the carrier and is only required to reach at the "modulation-effective status level" before the modulator. It signifies that the power or voltage of the input digital signal to the modulator could be low, as long as the input voltage or power reaches at a level by which the carrier can be effectively modulated. In this case, the digital signal is transported or manipulated between the local circuit blocks and is not required to be "power" transported. However, the modulated carrier after the modulator must be power-magnified and delivered to the antenna so that the modulated carrier is powerful enough to be propagated to a receiver located from the transmitter with a long distance. In the receiver, the modulated RF carrier can be de-modulated only if its power is strong enough to suppress the noise power at the input of the de-modulator. Typically, the ratio of RF signal to noise power at the input of the de-modulator is required to be more than 10 *dB*. It is therefore required that the RF signal must be "power" transported or operated before the de-modulator. After de-modulator the digital-type message is de-modulated from RF to base band. The digital signal is not required to be "power" transported but is required only to be "status" transported between local blocks for digital signal processing.

Summarily, the power-transport type of an RF signal and the status-transport type of a digital signal result from the actual engineering design requirement of the modulator and de-modulator when the digital data rate is low.

It is well known that the power of a signal delivered to a block or a part with an impedance Z can be expressed by

$$P = vi = \frac{v^2}{Z} \ , \qquad (1.6)$$

where P = Power delivered to a block or a part,
v = Voltage across the block or part,
i = Current flowing through the block or part,
Z = Impedance of the block or part.

For a given value of power, v^2 is proportional to Z. It implies that a higher voltage must be built across the block with higher impedance, or, in other words, the voltage across the block can be lowered when the load impedance is reduced. From the viewpoint of either cost or engineering design of the circuit, the application of low voltage is much better than that of high voltage. This is why the input and output impedance in the RF blocks are intentionally assigned to be low, because only a lower voltage is needed to build the same power on a low impedance block.

However, it is just the opposite for a digital signal, which is asked to do status-transportation. For a given power, a higher impedance can have a higher voltage swing across a block, and then the signal can ON/OFF the device more effectively. Or, a lower power is enough to produce a given voltage swing over a block with a higher impedance.

Consequently, when a digital circuit is operating with a low data rate, the philosophy of both RF and digital circuit designs is significantly different. RF design engineers worry about impedance matching whereas digital design engineers are indifferent on it. In circuit simulation, RF design engineers prefer to work in the frequency domain while digital design engineers like to work in the time domain. Correspondingly, in the test laboratory, RF design engineers prefer to use spectrum or network analyzers while digital design engineers like to use oscilloscopes. When they sit down together and discuss the two kinds of circuit designs, it seems as if they are two different kinds of aliens from different planets. For example, the RF design engineers like to use *dBW* or dB_m as the unit for measuring the output of a block or system while the digital design engineers insist on using *dBV*. Even in certain websites or in some publications, these two kinds of "aliens" argue with each other. Each tries to prove that their design philosophy is better than the others'.

As a matter of fact, these two design philosophies are both right. They are different because they are in charge of different tasks and goals. Eventually they don't conflict from each other. In the end everybody agrees that the propagation of radio waves must obey the Maxwell's equations and that the relationship of the current and the voltage in the circuitry must obey Ohm's law.

1.1.2 Case #2: Digital Circuits at High Data Rate

As the electronic industry moved forward rapidly in 1990's, the digital data rate in a transceiver was raised up from the order of Mbps to 100 Mbps and eventually, over 10 Gbps. The digital circuits are operated in a high data rate as described by the definition (1.2) or the expression (1.5). The examples can be found in many communication systems, such as, the Gigabit Ethernet Transceiver, the 10-Gbps Optical Transceiver, and so on. The difference between RF and digital circuit design can be summarized as shown in Table 1.2.

Table 1.2 The difference between RF and digital circuit when data rate is high.

Item	*RF module/RFIC*	*Digital circuit (High data rate)*
Impedance	**Low (50 *ohm*s typically)**	**Low to High**
Impedance matching	**Important**	**Important**
Current	**High (*mA*)**	**Low to High (*μA to mA*)**
Bandwidth	**Narrow or intermediate**	**Wide**
Transportation type	**Power (Watt)**	**Status (Voltage) & Power (Watt)**

In the case of high digital data rates, the tasks of both types of circuits are still kept unchanged: the RF circuit is still working for the **power** while digital circuit is still working for the **status** transportation or operation. However, the gap of the design philosophy for both RF and digital circuit disappears because in the case of high data rates, the effective **status** transportation or operation becomes possible only when it is an effective **power** transportation or operation. Furthermore, it must be emphasized that in the case of high digital data rates, the impedance matching becomes even more important for a digital circuit than for an RF circuit when the data rate is the same order as the RF or higher than the RF frequency. The reason is that the digital signal is usually a rectangular pulse while an RF signal is usual sinusoidal. The former contains a wideband spectrum while the latter contains a narrowband spectrum relatively.

An engineer who designs digital circuits with a high data rate must have RF design experience or background. He/she must take care of impedance matching in a very serious way. In a digital circuit block without impedance matching, the digital voltage level would suffer from an additional attenuation, an additional jitter, an additional cross-talk, and eventually an additional bit error. He/she must take care of the layout work very carefully. The layout for a digital circuit with a low data rate is not too important: correct connections are all that need to be taken care of. The runners could be lined up in parallel so that the entire layout looks nice and neat. However, the layout for a digital circuit with a high data rate must be taken care of as seriously as for an RF circuit. Just like an RF circuit, a runner in a digital circuit with a high data rate contributes to cross-talk like a regular electronic component or part in the circuit operation. Runners which are put in parallel could cause additional mutual inductances, capacitances, and

eventually cross-talk.

The impedance of digital circuit blocks might be either low or high, depending on the function and the topology of the circuitry. For the sake of the power saving, most digital circuit blocks should have very high impedance. However, a digital circuit block with a low input or output impedance is sometimes necessary for impedance matching.

In order to ensure a high speed transportation or manipulation, the current drain in a digital circuit must be increased. The current drain might be up to the order of *mA* in the digital circuit blocks where the data rate is close to or higher than the frequency of an RF signal, though the current drain is still kept in the order of μA in most of digital circuit blocks where the data rate is much lower than the frequency of an RF signal.

1.2 Significance of Impedance Matching

1.2.1 Power Transportation from a Source to a Load

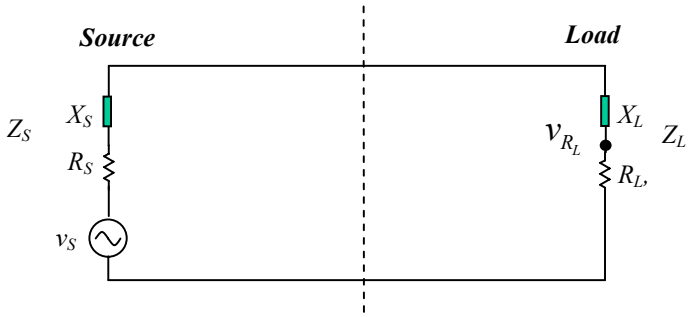


Figure 1.1 Voltage and power transported from a source to a load

Figure 1.1 depicts the voltage or power delivered from a source to a load, in which

$$Z_S = R_S + jX_S \ , \tag{1.7}$$

$$Z_L = R_L + jX_L \ , \tag{1.8}$$

where Z_S = Impedance of the source,
R_S = Resistor of the source,
X_S = Reactance of the source,
Z_L = Impedance of the load,
R_L = Resistor of the load,
X_L = Reactance of the load.

The X_s or X_L in the expression (1.7) or (1.8) is the reactance of either capacitor or inductor. It is well known that the average power across X_s or X_L over one period of the RF signal is zero. In other words, in the power transportation, the capacitor or inductor experiences only a process of charging and discharging but never receives any net power from the source. The power in the source can only be transported onto the load resistor, R_L. The power across the load resistor, R_L, can be expressed as

$$P_{R_L} = \frac{v_{R_L}^{\ 2}}{R_L} \ , \tag{1.9}$$

where P_{R_L} = Power transported from the source to the load resistor, R_L,

v_{R_L} = Voltage across the load resistor, R_L.

From Figure 1.1, we have

$$v_{R_L} = \frac{v_S}{Z_S + Z_L} R_L \ , \tag{1.10}$$

$$v_{R_L}{}^2 = \frac{v_S{}^2}{(R_S + R_L)^2 + (X_S + X_L)^2} R_L{}^2 \ . \tag{1.11}$$

From expressions (1.9) and (1.11), we have

$$P_{R_L} = v_S{}^2 \frac{R_L}{(R_S + R_L)^2 + (X_S + X_L)^2} \ . \tag{1.12}$$

Thus, it can be seen that the power transported from the source to the load resistor, P_{R_L}, depends not only on the value of R_L, but also on the value of R_S, X_S, and X_L.

1.2.2 Maximizing of Power Transportation without Phase Shift

From equation (1.12), it can be seen that the transported power, P_{R_L}, can reach a maximum when some specific relations between the source and the load impedance, Z_s or Z_L, are satisfied. Before searching for its maximum mathematically, one of the specific conditions to approach the maximum of P_{R_L} can be easily found from the denominator of the equation (1.12), that is,

$$X_S + X_L = 0 \ , \ \text{or,} \quad X_S = -X_L \ . \tag{1.13}$$

The relation (1.13) indicates that in order to achieve maximum power transportation from the source to the load resistor, the reactance of the source and the load must have equal magnitude but opposite sign. It implies that the load reactance, X_L, must be inductive if the source reactance, X_S, is capacitive, and vice versa. Under the condition of (1.13), equation (1.12) becomes

$$P_{R_L} = v_S{}^2 \frac{R_L}{(R_S + R_L)^2} \ . \tag{1.14}$$

Now let's find out another relation between the source and load impedances, Z_s and Z_L, for the maximum of transported power, P_{R_L}. By partially differentiating (1.14) in respect to R_L,

$$\frac{\partial P_{R_L}}{\partial R_L} = v_S{}^2 \left[\frac{1}{(R_S + R_L)^2} - 2\frac{R_L}{(R_S + R_L)^3} \right] = v_S{}^2 \frac{R_S - R_L}{(R_S + R_L)^3} \ . \tag{1.15}$$

The condition to maximize P_{R_L} is

$$\frac{\partial P_{R_L}}{\partial R_L} = 0 . \tag{1.16}$$

From relations (1.15), we have

$$R_S = R_L . \tag{1.17}$$

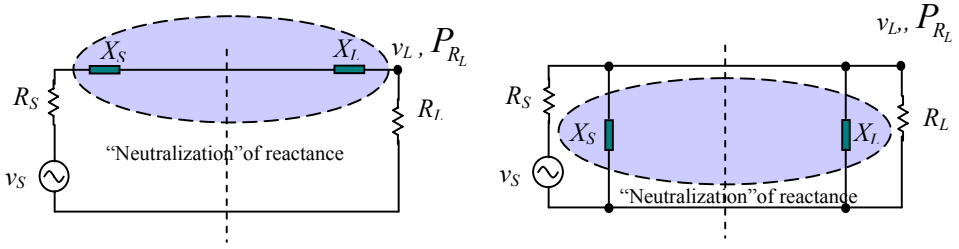
By combining of the relations, (1.13) and (1.17), it results

$$Z_S^* = Z_L , \quad \text{or} \quad Z_S^* = Z_L . \tag{1.18}$$

The relation (1.18) is the called the condition of conjugate impedance matching or simply the condition of impedance matching.

The condition of impedance conjugate matching not only enables maximum power transportation but also eliminates phase shift when power is transported from the source to the load. This is due to the fact that the resulting impedance contains only pure resistance in the entire source-load circuit loop under the condition of impedance conjugate matching. Further explanation can be conducted in terms of Figure 1.2(a) or 1.2(b). X_S and X_L are in series in Figure 1.2(a) while in parallel in Figure 1.2(b). When X_S is in series with X_L as shown in Figure 1.2 (a), their resultant reactance is zero due to the resonance in series. When X_S is in parallel with X_L as shown in Figure 1.2(b), their resultant reactance is infinitive due to the resonance in parallel. In both of cases, the reactances, X_S and X_L, are "neutralized" with each other and therefore, the phase shift of the voltage at the load from the voltage at the source is zero, that is,

$$\angle v_L - \angle v_S = 0 , \text{ or } \quad \angle v_L = \angle v_S . \tag{1.19}$$



(a) R_S in series with X_S
R_L in series with X_L

(b) R_S in parallel with X_S
R_L in parallel with X_L

Figure 1.2 Two matching cases when reactance of source is "neutralized" by reactance of load, or, vice versa, that is, $X_S = -X_L$

The zero phase shift of voltage is another important concept in impedance matching for the power transportation of an RF signal, especially for those circuits where the phase

modulation or frequency modulation is involved. Unfortunately it is often ignored or simply "swallowed" by the single concept of maximizing power transportation in the discussion about the impedance conjugate matching.

Under the condition of impedance conjugate matching, the maximum of power at the load is

$$P_{R_L} = P_{R_L,\max} = \frac{v_S^{\,2}}{4R_L} = \frac{v_S^{\,2}}{4R_S} \quad . \tag{1.20}$$

1.2.3 Conjugate Impedance Matching and Voltage Reflection Coefficient

In the cases of conjugate impedance matching between the source and the load, voltage reflection should not exist either at the source or at the load, that is, voltage reflection coefficient at the source or at the load, Γ_S and Γ_L, should be zero.

$$\Gamma_S = \frac{Z_S - Z_L}{Z_S + Z_L} = 0 \quad , \tag{1.21}$$

$$\Gamma_L = \frac{Z_L - Z_S}{Z_L + Z_S} = 0 \quad . \tag{1.22}$$

Both of expression (1.21) and (1.22) indicate that

$$Z_S = Z_L \ . \tag{1.23}$$

The expression (1.23) is apparently different from the condition of conjugate impedance matching (1.18).

As a matter of fact, the expression (1.23) is condition of maximum power transportation between the source, Z_S, and the load, Z_L, as shown in Figure (1.1) if

1) The source impedance $Z_S=R_S+jX_S$ is an indivisible entity or variable mathematically. In other words, *Xs* is always accompanied with *Rs* together and is not "neutralized" by any means.
2) The load impedance $Z_L=R_L+jX_L$ is an indivisible entity or variable mathematically. In other words, X_L is always accompany with R_L together and is not "neutralized" by any means.
3) Consequently, instead of P_{R_L} as shown in expression (1.9), the power delivered from source to load P_{Z_L} is concerned, that is,

$$P_{Z_L} = \frac{v_{Z_L}^{\ 2}}{Z_L} = v_S^{\ 2} \frac{Z_L}{(Z_S + Z_L)^2} \ . \tag{1.24}$$

In order to obtain the maximum of transported power, P_{Z_L}, let's partially differentiate (1.24) in respect to Z_L,

$$\frac{\partial P_{Z_L}}{\partial Z_L} = v_S^{\ 2} \left[\frac{1}{(Z_S + Z_L)^2} - 2\frac{Z_L}{(Z_S + Z_L)^3} \right] = v_S^{\ 2} \frac{Z_S - Z_L}{(Z_S + Z_L)^3} \tag{1.25}$$

The condition to maximize P_{Z_L} is

$$\frac{\partial P_{Z_L}}{\partial Z_L} = 0 . \tag{1.26}$$

From expression (1.25) and (1.26), it is therefore found that the expression (1.23) is the maximum condition of P_{Z_L} in respect to Z_L.

As a matter of fact, the conjugate impedance matching is a special case in the power transportation, in which the reactance of source and the reactance of load are "neutralized" from each other. Instead of power delivered from source Z_S to load Z_L, the power delivered from source R_S to load resistor R_L is concerned. On the other hand, instead of $Z_S = Z_L$ as shown in the expression (1.23), the condition of zero voltage reflection is reflected as $R_S = R_L$ as shown in the expression (1.17) since X_S and X_L are "neutralized" from each other.

It is therefore concluded that the condition of conjugate impedance matching (1.18) does not conflict with the expression (1.23) which represents the case of zero voltage reflection.

1.2.3 Impedance Matching Network

Usually the impedance of a source does not conjugate match with the impedance of the load, that is,

$$Z_S \neq Z_L^{\ *} \ . \tag{1.27}$$

In order to maximize power transportation without phase shift from the source to the load, the impedance conjugate matching condition must be satisfied. Therefore, an impedance matching network must be inserted between the source and the load. As shown in Figure 1.3, the input impedance of the impedance matching network must be equal to $Z_S^{\ *}$ and the

output impedance of the impedance matching network must be equal to Z_L^*, that is,

$$Z_{in} = Z_S^* \quad , \tag{1.28}$$

$$Z_{out} = Z_L^* \quad . \tag{1.29}$$

An impedance matching network in fact is an impedance-conversion network. It can be constructed by passive parts or both active and passive parts. The input and output matching networks of a LNA are a kind of matching networks which usually consists of only passive parts, i.e., capacitors and inductors. There would be no power consumption in the matching network if a matching network consists of only ideal inductors and capacitors. Consequently, up to Figure 1.3, we have

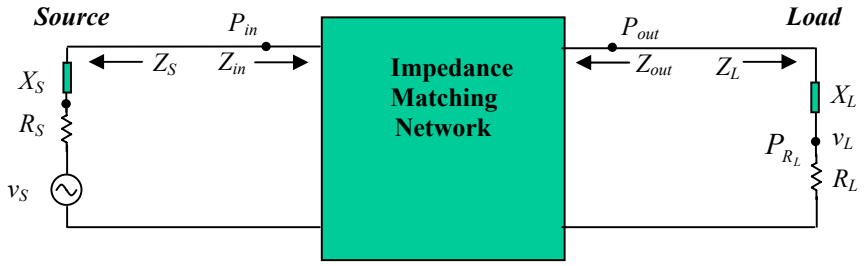


Figure 1.3 An impedance matching network is inserted between source and load when $Z_S \neq Z_L^*$

$$\frac{v_s^{\,2}}{4R_S} = \frac{v_s^{\,2}}{4R_{in}} = \frac{v_L^{\,2}}{4R_{out}} = \frac{v_L^{\,2}}{4R_L} \quad , \tag{1.30}$$

Or,

$$P_{R_S} = P_{in} = P_{out} = P_{R_L} \quad , \tag{1.31}$$

$$P_{R_S} = \frac{v_s^{\,2}}{4R_S}, \tag{1.32}$$

$$P_{in} = \frac{v_s^{\,2}}{4R_{in}} \quad , \tag{1.33}$$

$$P_{out} = \frac{v_L^{\,2}}{4R_{out}} \quad , \tag{1.34}$$

$$P_{R_L} = \frac{v_L^{\,2}}{4R_L}, \tag{1.35}$$

where R_{in} = real part of Z_{in},
R_{out} = real part of Z_{out} ,

P_{R_S} = power in the source,
P_{in} = power delivered into matching network from source,
P_{out}= power arrived at the output of impedance matching network,
P_{R_L} = power delivered onto the load resistor from source.

The expression (1.30) looks quite simple, but it implies two important concepts:

1) The powers, P_{R_S} , P_{in}, P_{out}, and P_{R_L} , are equal to each other. It is important in the practical power measurement because the power, P_{R_S} , at the source can be measured as its equal value, P_{R_L} , at the load.
2) The insertion of an impedance matching network without power consumption will ensure maximum power transportation without phase shift of the voltage from the source to the load.

There are two types of matching network: passive and active matching network. Usually a passive matching network consists of capacitors and inductors. The resistor is not a good member because it brings about noise and power gain reduction. An emitter follower, a source follower and a buffer are examples of another type of matching network, which are constructed by active and passive parts and are classified as active matching networks.

A question might be raised up in respect to the matching network itself, that is, within the impedance matching network, the impedance between the parts are generally not matched to each other. Is it necessary to insert a "sub-impedance matching network" between the two parts since they are not impedance matched with each other?

The answer is simple: It is not necessary to do so if the matching network is a basic power transportation unit in which their parts are indivisible in the power transportation, but it is necessary to do so if the matching network consists of more than one basic power transportation unit. Let's furthermore illustrate it by examples.

The main parts of an emitter follower are the emitter resistor and the transistor. By looking into the emitter of the transistor, it can be seen that its impedance is usually not conjugate-matched to the impedance of the emitter resistor. It is, however, not necessary to insert a "sub-impedance matching network" between the emitter resistor and the transistor, because both of parts constitute a basic unit in the power transportation and are indivisible. On the contrary, in the CE-CB cascode LNA design, a matching network might be inserted between the collector of CE transistor and the emitter of CB transistor because the 1st unit, composed by the CE transistor and the parts connected to its emitter and base, and the 2nd unit, composed by the CB transistor and the parts connected to its collector, are independent units in the power transportation, respectively. Very often, a matching network is not inserted between the collector of CE transistor and the emitter of CB transistor if the designer intentionally treats both of CE and CB transistors together as one power transportation unit. It, of course, results in a certain loss of power though it simplifies the circuit topology.

1.3 Problems due to Unmatched Status of Impedance

Impedance matching is important in RF circuits because it maximizes power transportation without phase shift of the voltage from the source to the load. Un-matched status of impedance could bring about a series of serious problems. We are going to discuss them in this section.

Un-matched status of impedance implies that the voltage or power reflection is happened at the source or at the load. In transmission line theory, the voltage reflection coefficients Γ_S and Γ_L have been defined and widely applied in the application of Smith Chart. For the simplicity in the discussion of the power transportation, the power reflection coefficient of the source and the load, γ_S and γ_L, is defined as

$$\gamma_S = \frac{P_{Z_S,reflected}}{P_{Z_S,insident}} = \Gamma_S{}^2 \ , \tag{1.36}$$

$$\gamma_L = \frac{P_{Z_L,reflected}}{P_{Z_L,insident}} = \Gamma_L{}^2 \ . \tag{1.37}$$

where γ_S = Power reflection coefficient at source,
γ_L = Power reflection coefficient at load,
Γ_S = Voltage reflection coefficient at source,
Γ_L = Voltage reflection coefficient at load,
$P_{Z_S,incident}$ = Incident power at source,
$P_{Z_S,reflected}$ = Reflected power at source,
$P_{Z_L,incident}$ = Incident power at load,
$P_{Z_L,reflected}$ = Reflected power at load.

1.3.1 General Expression of Power Transportation

As a matter of fact, equation (1.12) is an ideal description of power transportation for the matched condition, that is, in the cases when the power reflection coefficients of the source and the load, γ_S and γ_L, both equal zero.

In unmatched impedance cases, the power reflection coefficients of the source and the load, γ_S and γ_L, are not zero. Note that the voltage reflection coefficients of the load and the source are correlated by

$$\Gamma_L = \frac{Z_L - Z_S}{Z_L + Z_S} = -\Gamma_S \ , \tag{1.38}$$

then, from the definition (1.36) and (1.37), the power reflection coefficients, γ_S and γ_L, are the same and can be commonly denoted by γ, that is,

$$\gamma = \gamma_S = \gamma_L \, . \tag{1.39}$$

Figure 1.4 shows the unmatched cases when $\gamma \neq 0$.

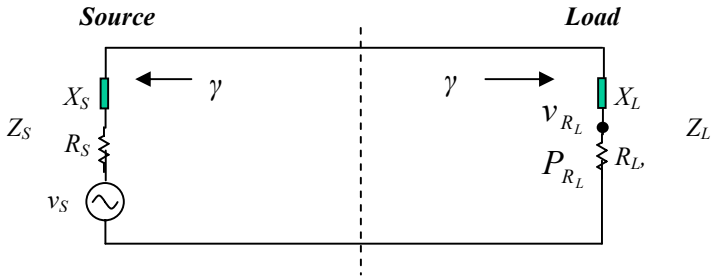


Figure 1.4 Voltage and power transported from a source to a load when $\gamma \neq 0$.

After the first reflection from the load, equation (1.12) must be modified to

$$P_{R_L} = v_s^{\,2} \frac{R_L}{(R_S + R_L)^2 + (X_S + X_L)^2} (1 - \gamma) \, ,$$

or,

$$P_{R_L} = \left(v_{so} \, e^{\,j\omega (t - t_d)}\right)^2 \frac{R_L}{\left|Z_S + Z_L\right|^2} (1 - \gamma) \ , \tag{1.40}$$

if the source,

$$v_S = v_{So} \, e^{\,j\omega t} \ , \tag{1.41}$$

where t_d = Delay time of power transportation from the source to the load.
The factor (1-γ) represents the portion of the coming power remaining on the load after the first reflection from the load.

The general expression of power transportation is a sum of all the reflected powers between source and load and can be expressed as follows:

$$P_{R_L}(t) \approx {v_{So}}^2 e^{j2\omega(t-t_d)} \frac{R_L}{\left|Z_S + Z_L\right|^2}(1-\gamma) + \\ + {v_{So}}^2 e^{j2\omega(t-3t_d)} \frac{R_L}{\left|Z_S + Z_L\right|^2}(1-\gamma)(\gamma)^2 + \\ + {v_{So}}^2 e^{j2\omega(t-5t_d)} \frac{R_L}{\left|Z_S + Z_L\right|^2}(1-\gamma)(\gamma)^4 + \\ + {v_{So}}^2 e^{j2\omega(t-7t_d)} \frac{R_L}{\left|Z_S + Z_L\right|^2}(1-\gamma)(\gamma)^6 + \\ + {v_{So}}^2 e^{j2\omega(t-9t_d)} \frac{R_L}{\left|Z_S + Z_L\right|^2}(1-\gamma)(\gamma)^8 + \ldots \qquad (1.42)$$

The second term on the right side of expression (1.42) represents the remaining power on the load after the second reflection from the load. The additional factor, γ^2, represents the power reflected back and forth one time between source and load. The third term on the right side of expression (1.42) represents the remaining power on the load after third reflection from the load. The additional factor, $(\gamma)^4$, represents the power reflected back and forth twice between source and load, and so on.

The expression (1.42) can be denoted by a summation symbol:

$$P_{R_L}(t) = {v_{So}}^2 \frac{R_L}{\left|Z_S + Z_L\right|^2}(1-\gamma)\sum_{n=0}^{\infty} e^{j2\omega[t-(2n+1)t_d]}(\gamma)^{2n} \quad . \qquad (1.43)$$

If

$$t_d \to 0 \quad , \qquad (1.44)$$

then

$$P_{R_L}(t) = {v_{So}}^2 e^{j2\omega t} \frac{R_L}{\left|Z_S + Z_L\right|^2} \frac{1}{1+\gamma} , \qquad (1.45)$$

Or,

$$P_{R_L}(t) = P_{R_L,\gamma=0}(t) \frac{1}{1+\gamma} \quad , \qquad (1.46)$$

where

$$P_{R_L,\gamma=0}(t) = {v_{So}}^2 e^{j2\omega t} \frac{R_L}{\left|Z_S + Z_L\right|^2} \ , \qquad (1.47)$$

which is the special $P_{R_L}(t)$ when $\gamma = 0$.

1.3.2 Power Instability and Additional Power Loss

In the case when the impedance is un-matched, strictly speaking, the power transported from the source to the load, P_{R_L}, is a time-variant, including both of its magnitude and phase. It can be found from expression (1.43) as long as t_d is taken account. The power at the load is varied from time to time so that it acts as the power instability. However, in most cases, especially in integrated circuit, the t_d usually is very short and such a power instability cannot be conceived because the length of the runner from the source to the load is usually much less than the corresponding quarter wavelength. Consequently such a power instability could be ignored and expression (1.45) is a good approximation.

In expression (1.45), the factor $R_L/|Z_S+Z_L|^2$ is the same as shown in the expression (1.12), which reaches its maximum, $1/(4R_L)$, when the impedances of the source and the load, Z_S and Z_L, are conjugate matched to each other. In unmatched cases it is smaller than $1/(4R_L)$. There is another additional power attenuation factor, $1/(1+\gamma)$, appearing in expression (1.45). It represents an additional power loss due to the un-matched status and results from the multi-reflections between the source and the load. It returns back to the matched case if $\gamma = 0$. In un-matched cases, the additional power attenuation factor, $1/(1+\gamma)$, is always less than 1 if $\gamma>0$.

Let's furthermore examine the additional power loss.

From expressions (1.46) we have

$$\Delta P_{R_L}(t) = P_{R_L}(t) - P_{R_L,\gamma=0}(t) = -P_{R_L,\gamma=0}(t)\frac{\gamma}{1+\gamma} \quad , \tag{1.48}$$

or,

$$\frac{\Delta P_{R_L}(t)}{P_{R_L,\gamma=0}(t)} = -\frac{\gamma}{1+\gamma} \quad . \tag{1.49}$$

As examples, Table 1.3 lists some of calculated additional power loss due to the unmatched cases in terms of expressions (1.48) and (1.49). The first row shows the cases when $\gamma = 0$, in which there is not additional power loss. The 2nd and 3rd row show the cases when γ is not zero but less than or equal to 10%, in which the additional power loss is a small amount. It is less than 0.5 *dB* from the original $P_{R_L,\gamma=0}$ =-30 dB_m. However, the additional power loss becomes an appreciable amount when γ >50%, in which the additional power loss is more than 1 *dB*. The additional power loss due to the un-matched circuit design could seriously damage the performance of a communication or other system. For instance, a communication system with 64 QAM modulation would require power accuracy between channels to be less than a couple of tenth dB. The unmatched design for the RF circuit might be a serious killer in a dark corner.

Table 1.3 Additional power loss due to the unmatched case when $P_{R_L,\gamma=0}$ =-30 *dBm*.

$$\Delta P_{R_L}(t) = -P_{R_L,\gamma=0}(t)\frac{\gamma}{1+\gamma}$$

$$P_{R_L}(t) = P_{R_L,\gamma=0}(t)\frac{\gamma}{1+\gamma}$$

γ, %	$\Delta P_{R_L}/P_{R_L,\gamma=0}$,%	$P_{R_L,\gamma=0}$,dBm	ΔP_{R_L},dBm	P_{R_L},dBm
0	0.00	**-30.00**	**-Infinitive**	**-30.00**
5	4.76	**-30.00**	**-43.22**	**-30.21**
10	9.09	**-30.00**	**-40.41**	**-30.41**
50	33.33	**-30.00**	**-34.77**	**-31.76**

1.3.3 Additional Distortion and Quasi-Noise

In unmatched cases, the sequentially reflected signal between the source and the load is added to the incoming signal and consequently disturbs the incoming signal either at the source or at the load sequentially. The disturbance can be catalogued into two types: additional distortion and quasi-noise. It brings about the additional distortion to the incoming signal when the frequency of signal is constant or when the sequential reflected signal has the same frequency as that of the incoming signal. Another kind of disturbance can be reluctantly named as "quasi-noise", in which the frequency of the signal is varied from time to time, or the frequency of the sequential reflected signal is different from the incoming one. It should be noted that the additional distortion due to unmatched impedance is not a stable spurious distortion, and the quasi-noise due to unmatched impedance is not a white noise.

Assuming that the powers at the load, $P_{R_{L1}}$ and $P_{R_{L2}}$, are observed at two instants, t and $t+2t_d$, respectively, which can be described in terms of equation (1.42):

$$P_{R_L1}(t) \approx {v_{S1o}}^2 e^{j2\omega_1(t-t_d)} \frac{R_L}{\left|Z_S + Z_L\right|^2}(1-\gamma) + \\ + {v_{S1o}}^2 e^{j2\omega_1(t-3t_d)} \frac{R_L}{\left|Z_S + Z_L\right|^2}(1-\gamma)(\gamma)^2 + \\ + {v_{S1o}}^2 e^{j2\omega_1(t-5t_d)} \frac{R_L}{\left|Z_S + Z_L\right|^2}(1-\gamma)(\gamma)^4 + \\ + {v_{S1o}}^2 e^{j2\omega_1(t-7t_d)} \frac{R_L}{\left|Z_S + Z_L\right|^2}(1-\gamma)(\gamma)^6 + \\ + {v_{S1o}}^2 e^{j2\omega_1(t-9t_d)} \frac{R_L}{\left|Z_S + Z_L\right|^2}(1-\gamma)(\gamma)^8 + ... \qquad (1.50)$$

$$P_{R_L2}(t+2t_d) \approx {v_{S2o}}^2 e^{j2\omega_2(t+t_d)} \frac{R_L}{\left|Z_S + Z_L\right|^2}(1-\gamma) + \\ + {v_{S2o}}^2 e^{j2\omega_2(t-t_d)} \frac{R_L}{\left|Z_S + Z_L\right|^2}(1-\gamma)(\gamma)^2 + \\ + {v_{S2o}}^2 e^{j2\omega_2(t-3t_d)} \frac{R_L}{\left|Z_S + Z_L\right|^2}(1-\gamma)(\gamma)^4 + \\ + {v_{S2o}}^2 e^{j2\omega_2(t-5t_d)} \frac{R_L}{\left|Z_S + Z_L\right|^2}(1-\gamma)(\gamma)^6 + \\ + {v_{S2o}}^2 e^{j2\omega_2(t-7t_d)} \frac{R_L}{\left|Z_S + Z_L\right|^2}(1-\gamma)(\gamma)^8 + ... \qquad (1.51)$$

Let's discuss the additional distortion first. Note that the frequency ω_1 in this case is the same as the frequency ω_2 in expression (1.50) and (1.51) and both of them can be commonly denoted by ω.

At the instant (t-t_d), the second term of $P_{R_{L2}}(t+2t_d)$ arrives at the load together with the first term of $P_{R_{L1}}(t)$. Assuming that the first term of $P_{R_{L1}}(t)$ is the desired signal power, that is,

$$S_{R_L1}(t-t_d) = {v_{S1o}}^2 e^{j2\omega(t-t_d)} \frac{R_L}{\left|Z_S + Z_L\right|^2}(1-\gamma) \ , \qquad (1.52)$$

then, the second term of $P_{R_{L2}}(t+2t_d)$ becomes the additional distortion to the $P_{R_{L1}}(t)$,

$$\Delta P_{R_L1}(t-t_d) = {v_{S2o}}^2 e^{j2\omega(t-t_d)} \frac{R_L}{\left|Z_S + Z_L\right|^2}(1-\gamma)\gamma^2 \ , \qquad (1.53)$$

where $S_{R_L1}(t-t_d)$ = Desired signal power arrived at load at the instant t-t_d,
$\Delta P_{R_{L1}}(t-t_d)$ = Additional distortion power arrived at load at the moment t-t_d.

The additional distortion, ΔD_{R_L}, can be evaluated directly from expressions (1.52) and (1.53) as follows:

$$\Delta D_{R_L} = \sqrt{\frac{\Delta P_{R_L1}}{S_{R_L1}}} = \gamma \quad . \tag{1.54}$$

Let's assume that

$$v_{S1o} = v_{S2o} \quad , \tag{1.55}$$

the resulting distortion, D_{R_L}, therefore is

$$D_{R_L} = D_{R_Lo} + \Delta D_{R_L} = D_{R_Lo} + \gamma \quad , \tag{1.56}$$

Table 1.4 Evaluated additional distortion as the reflection coefficient γ is varied, for the cases of D_{R_Lo},= *10%.*

$$D_{R_L} = D_{R_Lo} + \gamma$$

γ, %	D_{R_Lo},%	Phase, deg.	D_{R_L},%	Phase, deg.
0	**10**	**5.73**	**10.00**	**5.73**
5	**10**	**5.73**	**10.05**	**5.76**
10	**10**	**5.73**	**10.10**	**5.79**
50	**10**	**5.73**	**10.50**	**6.02**

$$D_{R_Lo} = \sqrt{\frac{\Delta P_{R_L1o}}{S_{R_L1}}} \quad , \tag{1.57}$$

is the distortion in the case of $\gamma = 0$.

As an example, Table 1.4 lists some calculated additional distortion due to the unmatched impedance. It can be seen that there is no additional distortion when $\gamma = 0$. In this case, $D_{R_L} = D_{R_Lo}$. The additional distortion is negligible when $\gamma<10\%$. However, it becomes conceivable when $\gamma>10\%$.

Now let's discuss the additional "quasi-noise". Note that the frequency ω_1 in this case is different from the frequency ω_2 in expression (1.50) and (1.51).

At the instant (t-t_d), the second power term of $P_{R_{L2}}(t+2t_d)$ arrives at the load together with the first term of $P_{R_{L1}}(t)$. Assuming that ω_1 is the angular frequency of a desired signal, then the first term of $P_{R_{L1}}(t)$ is the desired power of signal, that is,

$$S_{R_L1}(t-t_d) = V_{S1o}{}^2 e^{j2\omega_1(t-t_d)} \frac{R_L}{\left|Z_S+Z_L\right|^2}(1-\gamma) \ . \tag{1.58}$$

The second power term of $P_{R_{L2}}(t+2t_d)$ becomes the additional quasi-noise in respect to $P_{R_{L1}}(t)$ because its frequency is different from that of the desired signal, that is,

$$\Delta N_{R_L1}(t-t_d) = V_{S2o}{}^2 e^{j2\omega_2(t-t_d)} \frac{R_L}{\left|Z_S+Z_L\right|^2}(1-\gamma)\gamma^2 \ , \tag{1.59}$$

where $S_{R_L1}(t-t_d)$ = Desired signal power arrived at the load at the instant t-t_d,

$\Delta N_{R_L1}(t-t_d)$ =Additional quasi-noise power arrived at the load at the instant t-t_d.

Note assumption (1.55) and neglect the phase difference between $e^{j2\omega_1(t-t_d)}$ and $e^{j2\omega_2(t-t_d)}$, the original ratio of signal to noise without additional quasi-noise, SNR_{R_Lo} is

$$SNR_{R_Lo} = \frac{S_{R_L1}}{N_{R_L1}} \ . \tag{1.60}$$

From expressions (1.55), (1.58), (1.59) and (1.60), we have

$$\frac{\Delta N_{R_L1}}{N_{R_L1}} = (\gamma)^2 SNR_{R_Lo} \ . \tag{1.61}$$

With the additional quasi-noise power due to un-matched case, the ratio of signal to noise becomes

$$SNR_{R_L} = \frac{S_{R_LI}}{N_{R_L1}+\Delta N_{R_L1}} = \frac{SNR_{R_Lo}}{1+\dfrac{\Delta N_{R_L1}}{N_{R_L1}}} = \frac{SNR_{R_Lo}}{1+(\gamma)^2 SNR_{R_Lo}} \ , \tag{1.62}$$

where SNR_{R_L} = resulting ratio of signal to noise with additional quasi-noise due to unmatched impedance. As examples, Table 1.5 lists some calculated ratio of signal to noise due to the unmatched impedance.

Table 1.5 Calculated ratio of signal to noise as the reflection coefficient, γ, is varied.

$$SNR_{R_L} = \frac{SNR_{R_L o}}{1 + (\gamma)^2 SNR_{R_L o}}$$

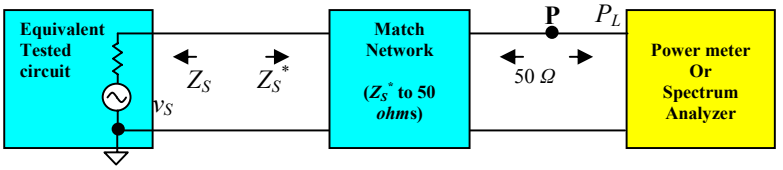
γ, %	$SNR_{R_L o}$,dB	$SNR_{R_L o}$,W	γ^2, %	$1+\gamma^2 SNR_{R_L o}$,W	$1+\gamma^2 SNR_{R_L o}$,dB	SNR_{R_L} ,dB
0	**10**	**10**	0.00	1.00	0.00	**10.00**
5	**10**	**10**	0.25	1.03	0.11	**9.89**
10	**10**	**10**	1.00	1.10	0.41	**9.59**
50	**10**	**10**	25.00	3.50	5.44	**4.56**

From Table 1.5 it can be seen that the "quasi-noise" disappears and the ratio of signal to noise is not degraded if either γ =0. This is due to that in these cases the power reflection back and forth between the source and the load does not exist. However, the degradation of the ratio of signal to noise becomes conceivable when γ is increased up to 10% or more, and eventually a huge degradation from 10 *dB* to 4.56 *dB* happens when both of γ_S and γ_L are increased up to 50%.

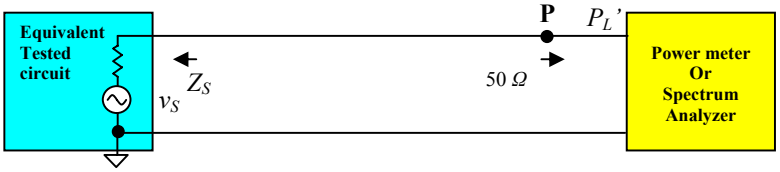
1.3.4 Power Measurement

Very often the power measurement is conducted by means of a power meter or a spectrum analyzer in a laboratory. The impedance of the spectrum analyzer is typically 50 Ω and the unit of power reading is usually *dB*. It is important to understand whether the test is under matched or unmatched conditions since the test outcomes are determined by the impedance matching status between the tested point and the spectrum analyzer. As an example, Figure 1.5 shows 2 cases of power test at point P. Assuming that the

impedance and the voltage of the equivalent tested circuit are Z_S and v_S respectively, and that the input impedance of the network analyzer is 50 Ω.



(a) Matched case



(b) Un-matched case

Figure 1.5 Output power of a tested block is measured by a power meter or a spectrum analyzer.

In the matched case (a), a matching network has been inserted between the tested circuit and the network analyzer. It has been discussed in section 1.2.3 and the power reading from the spectrum analyzer is

$$P_L = P_o = \frac{v_S^{\,2}}{4R_S} \quad , \tag{1.63}$$

where P_o= maximum of matched power, which can be sensed by the spectrum analyzer.

In the un-matched case (b), the spectrum analyzer directly measures the power at point P without any assistance, we have

$$P_L' = v_S^{\,2} \frac{50}{|50 + Z_S|^2} \quad , \tag{1.64}$$

$$\frac{P_L'}{P_o} = 4R_S \frac{50}{|50 + Z_S|^2} = \frac{200\,R_S}{(50 + R_S)^2 + X_S^{\,2}} \tag{1.65}$$

Table 1.6 lists the calculated results from expressions (1.65). It can be seen that the impedance matching condition is satisfied only when R_S= 50 Ω, X_S= 0 Ω. In such a special case, the impedance of the tested circuit is matched with that of the spectrum analyzer. It results that the measured power, P_L', is equal to the expected power, P_o, as shown in case (a). In the unmatched case (b), the measured power, P_L', will deviate from the expected power, P_o, with a certain amount. The deviation becomes significant when the impedance is far from matched condition. The readings of power measurement by a spectrum analyzer are reliable only when the testing is under the condition of impedance matching.

Table 1.6 Calculated values of P_L'/P_o in the un-matched case (b)

R_S	X_S	P_L'/P_o	P_L'/P_o, dB
10	0	0.555556	-2.6
10	50	0.327869	-4.8
10	100	0.147059	-8.3
10	1000	0.001993	-27.0
10	10000	0.000020	-47.0
50	**0**	**1.00**	**0.0**
50	50	0.800000	-1.0
50	100	0.500000	-3.0
50	1000	0.009901	-20.0
50	10000	0.000100	-40.0
100	0	0.888889	-0.5
100	50	0.800000	-1.0
100	100	0.615385	-2.1
100	1000	0.019560	-17.1
100	10000	0.000200	-37.0
1000	0	0.181406	-7.4
1000	50	0.180995	-7.4
1000	100	0.179775	-7.5
1000	1000	0.095125	-10.2
1000	10000	0.001978	-27.0
10000	0	0.019801	-17.0
10000	50	0.019801	-17.0
10000	100	0.019800	-17.0
10000	1000	0.019607	-17.1
10000	10000	0.009950	-20.0

1.3.5 Power Transportation and Voltage Transportation

Let's discuss a plausible question, that is, impedance matching is important to the power transportation, is it important to the voltage transportation? For instance, is it necessary to do impedance matching at the LO injection port when a mixer is designed?

Some engineers design mixers without impedance matching at the LO injection port. The reasons are

- As long as the runner from the LO source to the LO injection port in the mixer is short enough, the voltage from source would be directly effective on that port.

- Instead of power, only voltage is needed to ON/OFF the LO port of the transistor in a mixer.
- Or, alternatively, one can simply put a 50 Ω of resistor at the LO injection port so that it is "matching" to its 50 Ω source as shown in Figure 1.6.

The first reason is obviously wrong. As long as the impedance of the source is not matched with the impedance of the load, the power will be reflected back and forth between the source and the load, even if the length of runner approaches to zero.

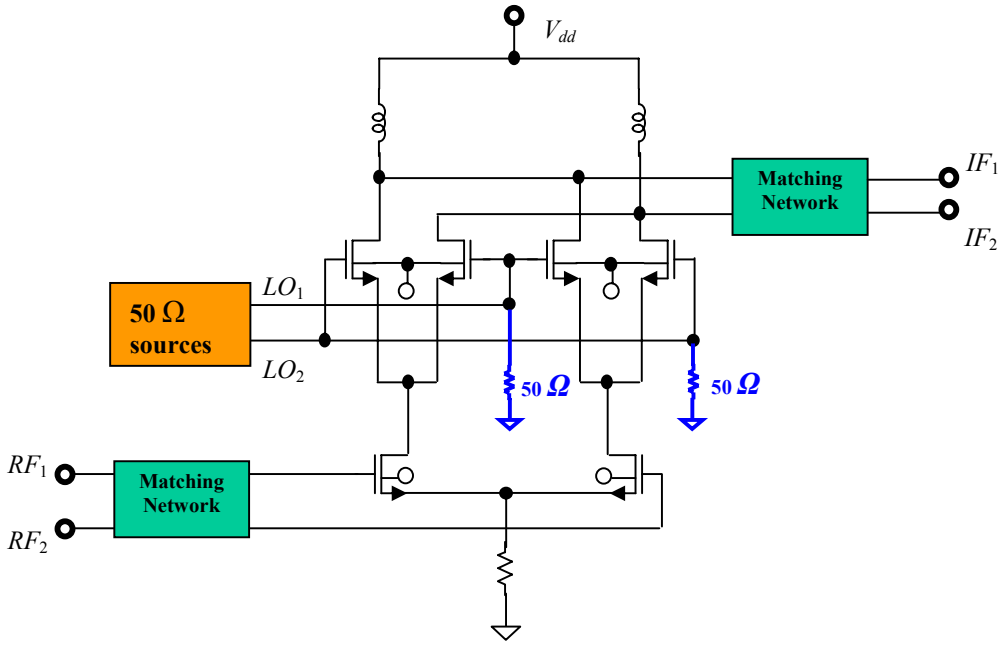


Figure 1.6 Incorrect impedance matching by attaching of two 50 Ω resistors at LO ports of a differential mixer.

It is true that only voltage is needed to ON/OFF the LO port of the transistor in a mixer. However, the voltage cannot be effectively transported from the LO injection source to LO port of the mixer if the impedance of the LO injection source is not matched to the impedance of the LO port. Let's analyze the voltage swing at LO port when the mixer as shown in Figure 1.7 is in 3 different cases:

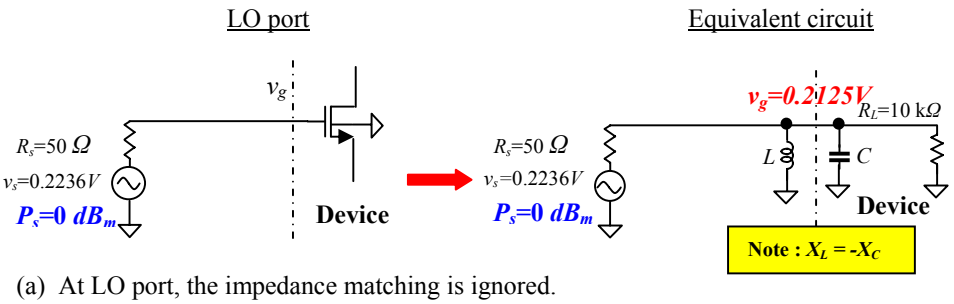
(a) At LO port, the impedance matching is ignored;
(b) At LO port, a 50 Ω resistor is connected in parallel;
(c) At LO port, impedance is well-matched by inserted a matching network between the LO injection source and the LO injection port.

Let's assume that

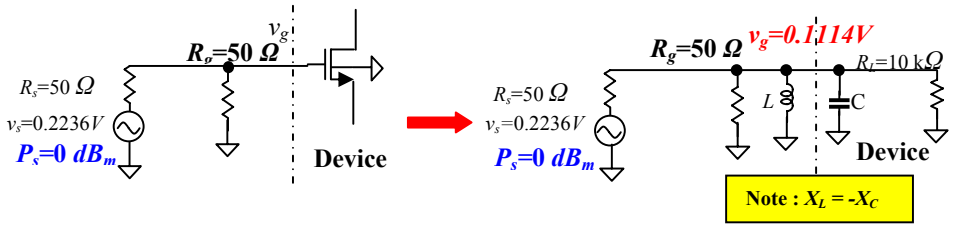
- The LO injection is a source with $v_s = 0.2236V$, or P_s=0 dB_m and R_s =50 Ω ;
- Looking into the LO injection port of the mixer, the impedance consists of a capacitor, C, and a 10 kΩ resistor, R_L, in parallel.

- The capacitor C is resonant with an inductor L, so that it is "neutralized", that is,

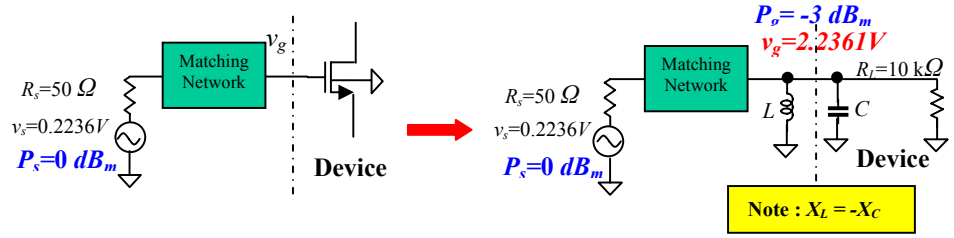
$$X_L = -X_C \quad . \tag{1.66}$$



(a) At LO port, the impedance matching is ignored.



(b) At LO port, a 50 Ω resistor is connected in parallel



(c) At LO port, impedance is well-matched by inserting of a matching network between the LO injection source and the LO injection port.

Figure 1.7 Three different ways for impedance matching at the LO port of a mixer

Consequently, the voltage at the LO port is

- $v_g = 0.2125V$ in case (a), in which the impedance has not been matched;
- $v_g = 0.1114V$ in case (b), in which a 50 Ω resistor has been connected to LO port in parallel;
- $v_g = 2.2360V$ in case (c), in which impedance has been well matched at the LO port.

In case (a), the impedance of LO port is not matched with its 50 Ω source. The gate voltage, $v_g = 0.2125V$, is dropped from $v_S = 0.2236V$ very little because the value of R_S, 50 Ω, is much lower than the value of R_L, 10 kΩ.

In case (b), the impedance of the LO port looks "well-matched" since the gate resistor, R_g = 50 Ω, is in parallel connected to the LO port, the gate of the device. As a matter of fact, it is matched to the power from the 50 Ω source to the gate resistor, R_g = 50 Ω, but not to the real part of impedance at the LO port, R_L=10 kΩ. In other words, the power is transported from the source to R_g but not to LO port. It results in the gate voltage, v_g = 0.1114V, which is dropped from v_S = 0.2236V by approximately 50% because R_g, 50 Ω, is connected to the gate of device, the LO port of the mixer. Consequently, it makes the voltage swing at the LO port about 50% less than that of an un-matched case (a).

In case (c), the impedance of the LO port is matched with its 50 Ω source by applying a matching network. The power is transported from the 50 Ω source to the output of matching network or the input of the LO port. The value of the power is about 3 dB lower than that in the source. That is, approximately, equal to – 3 dB_m since the power of the source is 0 dB_m. Consequently, the voltage at LO port can be pumped up to 2.2360V because

$$P_g = 10 \log \frac{v_g^{\,2}}{R_L} \quad , \tag{1.67}$$

$$-3(dBm) = 10 \log \frac{2.2360^{\,2}(\mathrm{V}^2)}{10000\ (\Omega)} \quad . \tag{1.68}$$

From the above analysis, it is therefore concluded that impedance matching is not only important to the power transportation, but also important to the voltage transportation. It is absolutely necessary to do impedance matching at the LO injection port, especially when a mixer with MOSFET devices is designed. Some engineers, who design a mixer without impedance matching at the LO injection port of a MOSFET device, are going down the wrong way.

From equation (1.68) it can be understood that for a fixed value of P_g, v_g is pumped up to a higher value as R_L is increased. In the case (c), the v_g is pumped up to 2.2360V because R_L is in a high value, 10 kΩ. Then, how about the case if the mixer is built by the bipolar transistor, where the LO port is the base of the transistor and its impedance is low or comparable with the 50 Ω source? Is it necessary to do impedance matching at the LO injection port when a mixer is built by bipolar transistors? The answer is yes again. In the case of a mixer built by bipolar transistor, equation (1.67) becomes

$$P_b = 10 \log \frac{v_b^{\,2}}{R_L} \quad , \tag{1.69}$$

where v_b = voltage at the base of the transistor or the LO port,
P_b = power at the base of the transistor or the LO port.
In order to obtain the maximum of voltage swing v_b at the base of the transistor or the LO port, the power must be effectively transported from the source to the LO port so that the P_b in equation (1.69) can reach to its maximum.

1.3.6 Burning of a Transistor

Finally, the most serious problem is the damage of a transistor if the unmatched condition exists at its input or output, especially in a high power amplifier design. As shown in Figure 1.8, assuming that the output port is not matched to the load, the reflected power could destroy the transistor. Usually an appropriate protection must be made in the test set-up for a power amplifier.

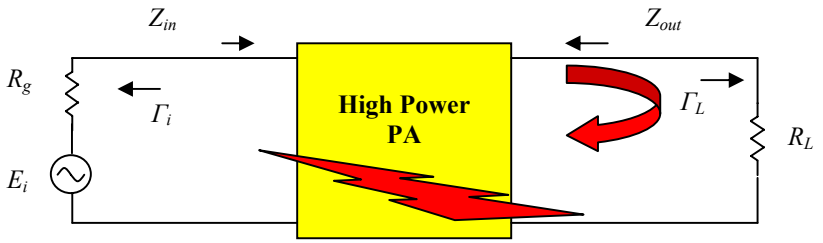


Figure 1.8 Possible burning of devices in the bench work for a PA.

References

[1] Youla, D. C., "A New Theory of Broad-Band Matching," IEEE Transactions on Circuit Theory, Vol. CT-11, pp30-50, March, 1964.

[2] ---, "S-Parameters, Circuit Analysis and Design," Hewlett-Packard Application Note 95, September, 1968.

[3] P. H. Smith, "Electronic Applications of the Smith Chart," (Book), McGraw-Hill, New York, 1969.

[4] B. S. Yarman, "New Approaches to Broadband Matching Problems," IMS Workshop, June 1983.

[5] B. S. Yarman, "Modern Approaches to Broadband Matching Problems," IEE Proceedings, Vol. 132, No. 2, pp.87-92, April, 1985.

[6] P.L.D. Abrie, "The Design of Impedance Matching Networks for Radio-Frequency and Microwave Amplifiers," (Book), Artech House, Norwood, Mass. 1985.

[7] J. D. Sifri, "Matching Technique Yields Optimum LNA Performance," Microwaves & RF, pp.87-90, February, 1986.

[9] J. –M. Collantes, R. D. Pollard, and M. Sayed, "Effects of DUT mismatch on the noise figure characterization: a comparative analysis of two Y-factor techniques," Instrumentation and Measurement, IEEE Transactions on , Vol. 51 , Issue: 6 , pp. 1150 – 1156, December 2002.

Index

Contents

Chapter 2 Impedance Matching

2.1 Impedance Measurement

2.1.1 Impedance Measured by S Parameter Measurement

In RF laboratories, the impedance of a basic part, such as a capacitor, an inductor, or a resistor, can be measured by using an impedance meter or a network analyzer. The impedance of a block, a sub-system, or an entire system can also be measured by using an impedance meter or a network analyzer. All of them are based on the principle of small signal measurement.

In a simulation, the impedance measurement of a DUT (Device Under Test) is usually executed by a network analyzer. The network analyzer measures the S parameter at one or two ports, and then the impedance is calculated from the S parameters or directly read from the Smith chart. Figure 2.1 shows its measuring principle by a network analyzer.

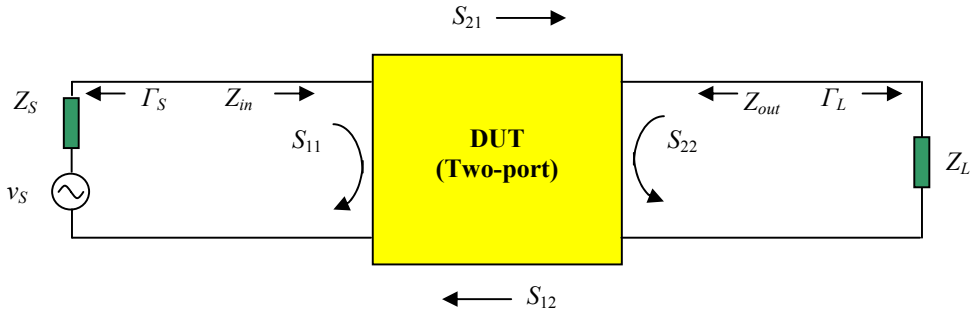


Figure 2.1 Principle of impedance measurement by a network analyzer

Theoretically the relations between input and output voltage reflection coefficients, Γ_{in} and Γ_{out}, and input and output S parameters, S_{11} and S_{22}, are

$$\Gamma_{in} = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L} , \tag{2.1}$$

$$\Gamma_{out} = S_{22} + \frac{S_{12}S_{21}\Gamma_S}{1 - S_{22}\Gamma_S} . \tag{2.2}$$

The Γ_{in}, Γ_{out} would be equal to S_{11}, S_{22} respectively, that is,

$$\Gamma_{in} = S_{11} , \quad \Gamma_{out} = S_{22} , \tag{2.3}$$

if

$$\Gamma_S = \Gamma_L = 0 , \tag{2.4}$$

or,

$$S_{12} = 0 , \tag{2.5}$$

or,

$$S_{21} = 0 \quad . \tag{2.6}$$

The condition (2.4) means that the cables, port 1 and port 2 are well-calibrated in the impedance measurement by network analyzer. The condition (2.5) or (2.6) means that the DUT is in an ideal state of either forward or backward isolation. Then we have

$$Z_{in} = \frac{1+\Gamma_{in}}{1-\Gamma_{in}} = \frac{1+S_{11}}{1-S_{11}} \quad , \tag{2.7}$$

$$Z_{out} = \frac{1+\Gamma_{out}}{1-\Gamma_{out}} = \frac{1+S_{22}}{1-S_{22}} \quad . \tag{2.8}$$

As shown in Figure 2.2, impedance measurement therefore can be conducted by means of the S_{11} or S_{22} testing with a network analyzer. Their values can be directly read out on the Smith Chart.

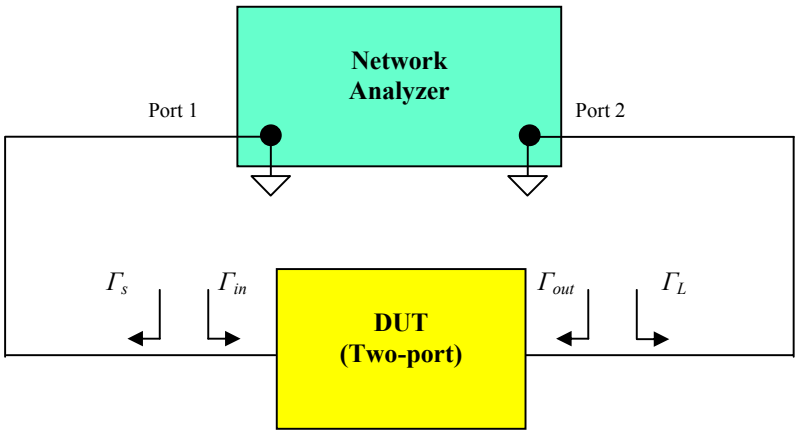


Figure 2.2 Impedance measurement by a network analyzer

2.1.2 Smith Chart: Impedance and Admittance Coordination

Smith Chart is the main tool in the impedance measurement and the impedance matching work. As a matter of fact, the Smith Chart is a representation of the reflection coefficient complex plane, that is,

$$\Gamma = U + jV, \tag{2.9}$$

In transmission line theory, the relationship between the reflection coefficient and the impedance is known as:

$$\Gamma = \frac{Z - Z_o}{Z + Z_o} \quad , \tag{2.10}$$

where Z_o is the characteristic impedance or reference impedance of the transmission line.

By the introduction of the normalized impedance, that is,

$$z = \frac{Z}{Z_o} \quad . \tag{2.11}$$

Expression (2.10) becomes

$$\Gamma = \frac{z-1}{z+1} = \frac{(r-1)+jx}{(r+1)+jx} \quad , \tag{2.12}$$

where r is normalized resistance and x is normalized reactance, that is,

$$z = r + jx \quad . \tag{2.13}$$

By comparing the real and imaginary parts of expressions (2.9) and (2.12), we have

$$U = \frac{r^2 - 1 + x^2}{(r+1)^2 + x^2} \quad , \tag{2.14}$$

$$V = \frac{2x}{(r+1)^2 + x^2} \quad . \tag{2.15}$$

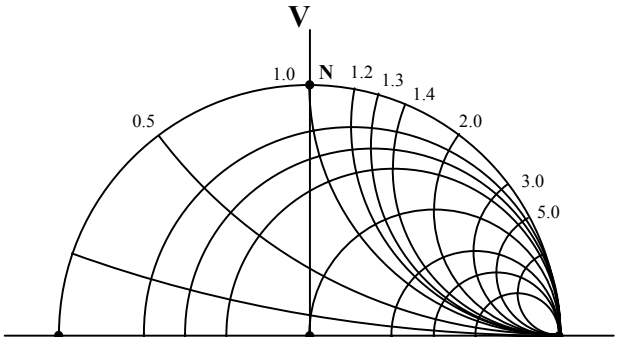
Eliminating x from (2.14) and (2.15), it results in

$$\left(U - \frac{r}{r+1}\right)^2 + V^2 = \left(\frac{1}{r+1}\right)^2 \quad . \tag{2.16}$$

This is the equation of a family of circles centered at $U = r/(r+1)$, $V=0$ with radii $1/(r+1)$. On the other hand, eliminating of r from (2.14) and (2.15), it results in

$$(U-1)^2 + \left(V - \frac{1}{x}\right)^2 = \left(\frac{1}{x}\right)^2 \quad . \tag{2.17}$$

This is the equation of a family of circles centered at $U = 1$, $V=1/x$ with radii $1/x$.



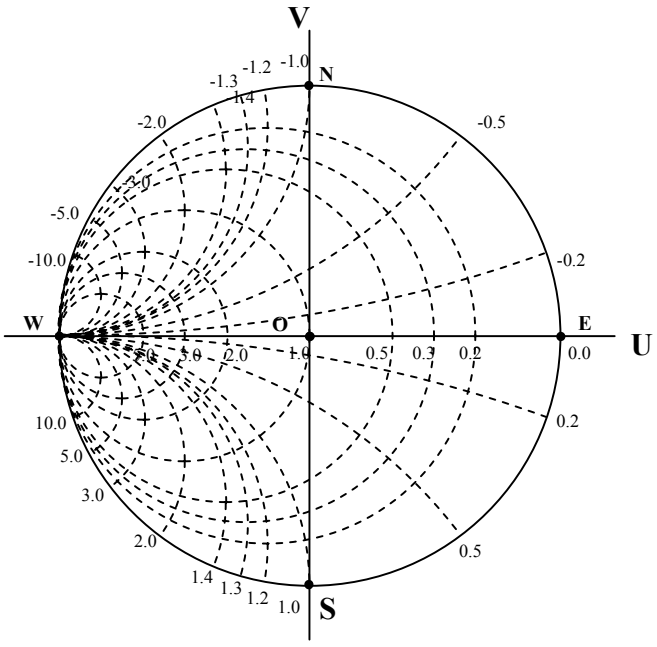
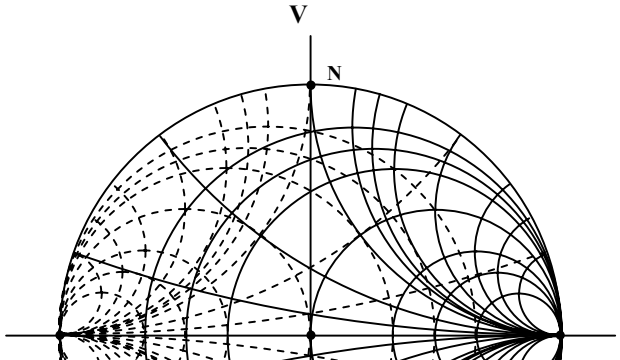


Figure 2.4 Admittance coordination of the Smith Chart



Sometimes it is more convenient to coordinate the Smith Chart by admittance, y, instead of impedance z,

$$y = \frac{1}{z} = g + jb \quad , \tag{2.18}$$

where y = normalized admittance,
g = normalized conductance,
b = normalized susceptance.

Figures 2.3 and 2.4 show the Smith Chart with impedance and admittance coordination respectively. As a matter of fact, Figure 2.4 can be obtained simply by $e^{j\pi}$ or 180^{o} rotating of Γ from Figure 2.3. From (2.12), we have

$$z = \frac{1+\Gamma}{1-\Gamma} \quad , \tag{2.19}$$

then,

$$y = \frac{1}{z} = \frac{1-\Gamma}{1+\Gamma} = \frac{1+\Gamma e^{j\pi}}{1-\Gamma e^{j\pi}} \quad . \tag{2.20}$$

It verifies that the conversion from z to y on the Smith Chart needs simply a 180^{o} rotation. In Figure 2.3, the numbers along the U axis denote the values of the normalized resistance r for the circles, and the numbers along the biggest circle denote the values of

the normalized reactance x arcs. In Figure 2.4, the numbers along the U axis denote the values of the normalized conductance g circles, and the numbers along the biggest circle denote the values of the normalized susceptance b arcs. Figure 2.5 is the combination of Figure 2.3 and 2.4, that is, a Smith Chart with both of impedance and admittance coordination, which is a powerful tool for impedance matching. If a matching network consists of only passive parts either in series or in parallel, then the impedance coordination is applied when the part is added in series, while the admittance coordination is applied when the part is added in parallel. Consequently, by the help of a Smith Chart as shown in Figure 2.5, the variation of impedance for a matching network can be easily figured out.

There are some special points in the Smith Chart. The center, denoted by O, is the point where r=1, x=0, or R=Z_o, X=0. It is a point where the impedance is equal to its characteristic impedance. The left-most point, denoted by W, is the point where r=0, x=0, or z=0. It is a point where the impedance is in a "short-circuited" state. The right-most point, denoted by E, is the point where x=0 but r is infinitive, or z is infinitive. It is the point where the impedance is in a "open-circuited" state. The upper-most point, denoted by N, is the point where r=0, x=1. It is the point where the impedance is purely inductive and X=Z_o. The bottom-most point, denoted by S, is the point where r=0, x=-1. It is the point where the impedance is purely capacitive and X=-Z_o.

In addition to both normalized impedance, z, and admittance, y, a Smith Chart is also coordinated for the voltage reflection coefficient, Γ, and other related parameters as listed as follows:

- Reflection coefficient of power, γ — Γ^2,
- Return loss in *dB*, — $-10log(\Gamma^2)$,
- Reflection loss in *dB*, — $-10log(1-\Gamma^2)$,
- *VSWR* (Voltage Standing Wave Ratio), — $(1+|\Gamma|)/(1-|\Gamma|)$,
- *VSWR* in *dB*, and — $20log(1+|\Gamma|)/(1-|\Gamma|)$,
- Transmission loss coefficient, — $(1+|\Gamma|^2)/(1-|\Gamma|^2)$.

The multiple coordinates make the Smith Chart more powerful in the engineering design. Most important parameters can be read directly from same chart without transformation from one to the other. In the LNA design, noise circle, gain circle, and stability circle can be developed and displayed on the Smith Chart.

2.1.3 Accuracy of Smith Chart

It should be noted that the accuracy of a reading from a Smith Chart depends on the location of the reading. Figure 2.6 shows the relatively inaccurate and accurate area on a Smith Chart.

In the areas where the impedance is not too low and not too high, the accuracy of impedance reading is acceptable and reliable. However, in the areas where the

impedance is very low or very high, the accuracy of the impedance reading might be questionable. For example, in the impedance measurement of a short whip antenna, which is operating around 27 MHz, the impedance might be directly displayed on the screen of a network analyzer via the Smith Chart, such as

- Real part: -1000 to +1500 Ω;
- Imaginary part: 3.0 pF.

The real part of an impedance is rapidly flashed between -1000 Ω to +1500 Ω and therefore it cannot be determined to a definite value. In such a case, an alternative test way must be found or another tool for the impedance measurement might be needed.

2.1.4 Relationship between Impedance in Series and in Parallel

The impedance reading from a test includes both real and imaginary parts, which are usually expressed in series as shown in expression (2.13). Sometimes the impedance with its real and imaginary part in series must be converted into the impedance with its real and imaginary part in parallel. Figure 2.7 sketches the real and imaginary parts of the impedance expressed in series (a) or in parallel (b). Their relations are

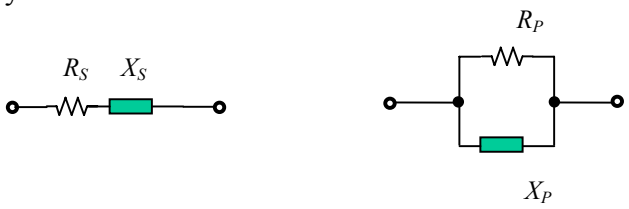
$$R_S + jX_S = R_P \,//\, jX_P = \frac{X_P{}^2 R_P + jX_P R_P{}^2}{R_P{}^2 + X_P{}^2} \quad , \qquad (2.21)$$

$$Q = \frac{|X_S|}{R_S} = \frac{R_P}{|X_P|} \quad , \qquad (2.22)$$

$$R_p = R_S\left(Q^2 + 1\right) \approx \frac{X_S{}^2}{R_S} \quad , \quad \text{if } Q \gg 1, \qquad (2.23)$$

$$X_p = X_S \frac{Q^2 + 1}{Q^2} \approx X_S \quad , \quad \text{if } Q \gg 1, \qquad (2.24)$$

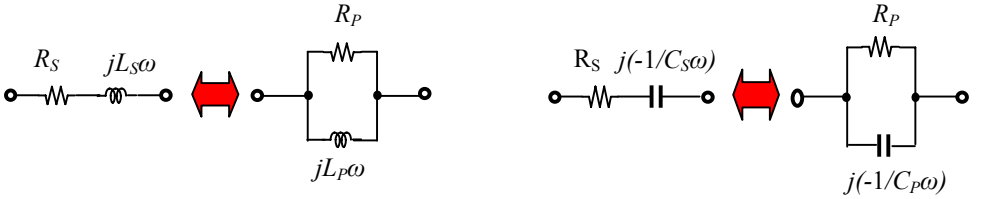
where Q = Quality factor .



(a) Impedance in series (b) Impedance in parallel

Figure 2.7 Real and imaginary parts of impedance expressed in series (a) or in parallel (b)

Figure 2.8 shows the conversion when the reactance is inductive and capacitive respectively.



$$X_S = X_{LS} = L_S\omega \qquad X_S = X_{CS} = -\frac{1}{C_S\omega}$$

$$X_P = X_{LP} = L_P\omega \qquad X_P = X_{CP} = -\frac{1}{C_P\omega}$$

(a) Reactance is inductive (b) Reactance is capacitive

Figure 2.8 Conversion of impedance between in series and in parallel

When the reactance is inductive, we have

$$R_S + jL_S\omega = R_P \,//\, jL_P\omega = \frac{{L_P}^2\omega^2 R_P + jL_P\omega {R_P}^2}{{R_P}^2 + {L_P}^2\omega^2}, \tag{2.25}$$

from (2.22), Q factor is

$$Q = \frac{L_S\omega}{R_S} = \frac{R_P}{L_P\omega} \quad , \tag{2.26}$$

$$R_p = R_S\left(Q^2 + 1\right) \quad , \tag{2.27}$$

$$L_p = L_S\,\frac{Q^2 + 1}{Q^2} \quad . \tag{2.28}$$

When the reactance is capacitive, we have

$$R_S + j\left(-\frac{1}{C_S\omega}\right) = R_P \,//\, j\left(-\frac{1}{C_p\omega}\right) = \frac{R_P - jC_P\omega {R_P}^2}{1 + {R_P}^2{C_P}^2\omega^2}, \tag{2.29}$$

from (2.22), Q factor is

$$Q = \frac{1}{R_S C_S\omega} = R_P C_P\omega \quad , \tag{2.30}$$

$$R_p = R_S\left(Q^2 + 1\right) \quad , \tag{2.31}$$

$$C_p = C_S\,\frac{Q^2}{Q^2 + 1} \quad . \tag{2.32}$$

2.2 Impedance Measured by Large Signal

So far, we have discussed about impedance measurement by means of S parameter measurement. It is well-known that the S parameter is correct only for a linear part, on which the testing signal is small. For the impedance measurement of a desired test unit, which is operating with high power or high voltage, the S parameter testing is no longer suitable. The impedance measurement in this case can be conducted by means of a circulator and vector voltmeter in a laboratory test.

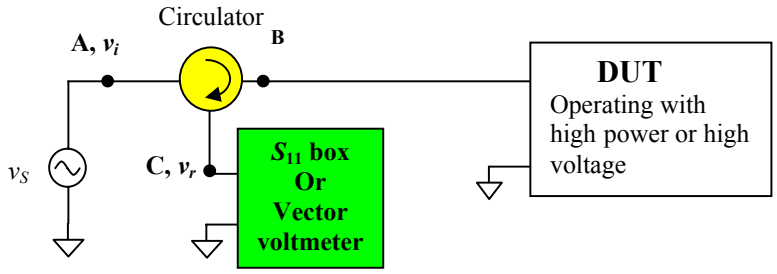


Figure 2.9 Impedance testing of a desired test unit with high power or high voltage

Figure 2.9 shows the setup for the impedance testing of a DUT which is operating with high power or high voltage. The special feature of a circulator is that the input power or voltage can be transported in only one direction, either clockwise or counter-clockwise. As shown in Figure 2.8, the input voltage or power at point A can reach point B as incident voltage or power for the DUT. The reflected voltage or power from DUT can only be transported to Point C and cannot be returned to point A so that it would not disturb the measurement at the input, point A. Consequently, at point C the reflected voltage or power from the DUT can be correctly measured by the S_{11} box or vector voltmeter. The impedance can be calculated from the incident voltage measured at point A, v_i, and the reflected voltage measured at point C, v_r, by the relationship,

$$z = \frac{1 + \Gamma}{1 - \Gamma} \quad , \tag{2.33}$$

where

$$\Gamma = \frac{v_r}{v_i} \quad . \tag{2.34}$$

In a practical power amplifier design, the power amplifier is operating with high power input and output. In a practical mixer design, its RF input and the IF output is usually treated as a low power port whereas the LO injection must be treated as a high power port. The test set-up for the impedance measurement of a mixer can be shown in Figure 2.10.

The mixer shown in Figure 2.10 is a differential one. The power sources, v_{RF1} and v_{RF2}, v_{LO1} and v_{LO2}, v_{IF1} and v_{IF2}, all are differential ones and must be set up with a phase difference of 180^o between source "1" and "2". Special attention must be paid to the IF

portions. The incident power from the power sources, v_{IF1} and v_{IF2}, would be reflected to the S_{11} box or vector voltmeter. We can calculate the impedance from the ratio of the reflected and incident power, or the reflection coefficient. However, there is another unexpected IF power, the product of the RF input and LO injection, which will also flow into the S_{11} box or vector voltmeter. In order to separate these two parts of power getting into the S_{11} box or vector voltmeter, there are two ways to do the IF impedance measurement:

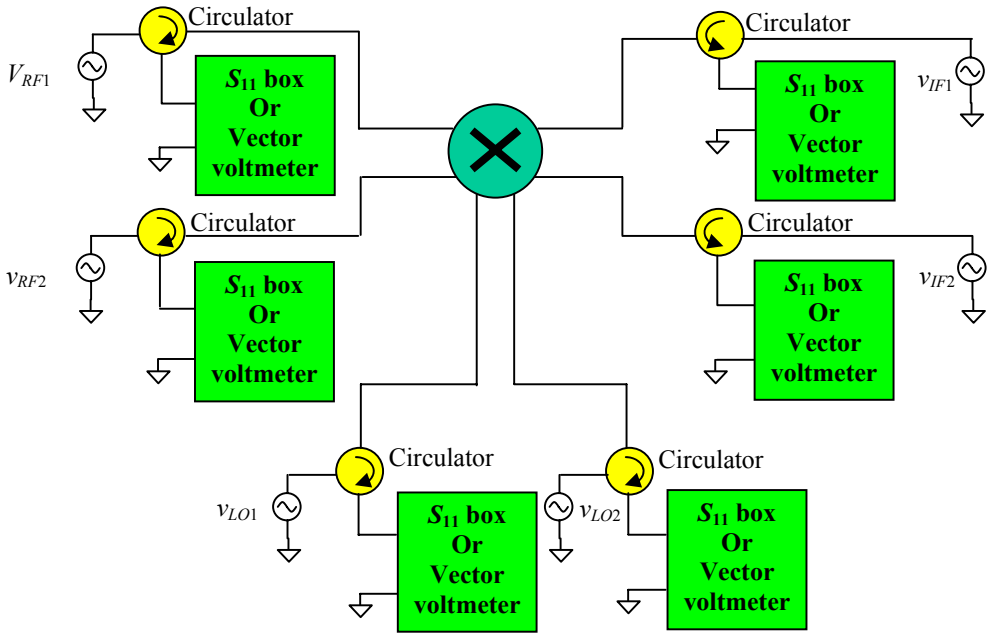


Figure 2.10 Impedance measurement set-up for a differential mixer

- The IF frequency at the IF power sources would be shifted a little bit so as to avoid the overlap with the product of RF input and LO injection. The measured impedance at IF ports, of course, corresponds to the shifted frequency but not to the expected IF frequency. Approximately we can treat the measured impedance at the shifted frequency as the IF impedance because the frequency shift is very small.
- Instead of the way of shifting the IF frequency, in the impedance testing for IF port, the IF product due to the RF input and LO injection can be prohibited if the RF input is turned-off.

It should be noted that another way to prohibit the IF product due to the RF input and LO injection might be to turn off the LO injection. This way is not allowed because the real operating status of a mixer is maintained by the LO injection.

To simplify the impedance testing at total 6 ports, 3 baluns can be applied so as to reduce the port number from 6 down to 3. Of course, the impedances of the baluns from single-ended to the differential must be well-known before they are applied to the test set-up. Figure 2.11 shows the set up to test a differential mixer with the application of baluns.

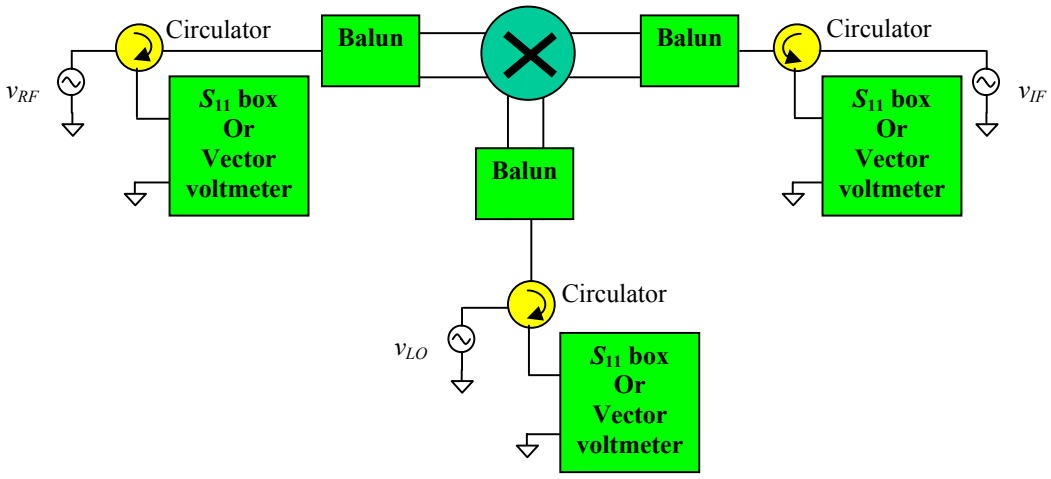


Figure 2.11 Impedance measurement set-up with baluns for a differential mixer

2.3 Impedance Matching

There are passive and active matching networks. In this section we are going to discuss passive matching network only.

A simple impedance matching can be done by a handy calculation through a couple of equations. However, by the assistance of Smith Chart, the impedance matching work becomes much easier. On the other hand, the source or load impedance is no more restricted to a pure resistor as treated in the previous section, but is usually a complex impedance, including both of real and imaginary parts on the Smith Chart.

There are many reference books about the applications of the Smith Chart in the engineering design. Here we are not attempting to repeat the description. However, we would like to re-emphasize one of the important applications, that is, to match a network to a known impedance, say, 50 Ω, which is always desired in the RF design. In order to pull impedance from somewhere to 50 Ω, a matching network should be added. Theoretically, most matching networks could be constructed by passive components such as inductors, capacitors, or resistors. The resistor usually is excluded because it attenuates the signal and introduces considerable noise. Almost nobody applies resistors in the matching work though it is allowed to be applied and is also depicted in Figure 2.12. Consequently, what we need to be familiar with is how to apply the capacitor and inductor into the matching network. Assuming that we are going to match one port to 50 Ω and start to add the matching parts one by one, let's examine how the impedance moves in the Smith Chart.

2.3.1 One Part Matching Network

Figure 2.12 shows the moving direction of the impedance at point P if one passive part, L, or C, or R, is added.

The variation of impedance on Smith Chart would obey the following thumb of rules if one inductor, or one capacitor, or one resistor is added to the original impedance P of the port:

- The addition of an inductor in series, L_S, results the original impedance P moving clockwise along the r = constant impedance circle. The moved arc length depends on the value of inductor;
- The addition of a capacitor in series, C_S, results the original impedance P moving counter-clockwise along the r = constant impedance circle. The moved arc length depends on the value of capacitor;
- The addition of an inductor in parallel, L_P, results the original impedance P moving counter-clockwise along the g = constant admittance circle. The moved arc length depends on the value of inductor;

- The addition of a capacitor in parallel, C_P, results the original impedance P moving clockwise along the g = constant admittance circle. The moved arc length depends on the value of capacitor.
- The addition of a resistor in series, R_S, results the original impedance P moving along the x = constant arc to a higher resistance circle. The moved distance depends on the value of resistor;
- The addition of a resistor in parallel, R_P, results the original impedance P moving along the x = constant arc to a lower resistance circle. The moved distance depends on the value of resistor.

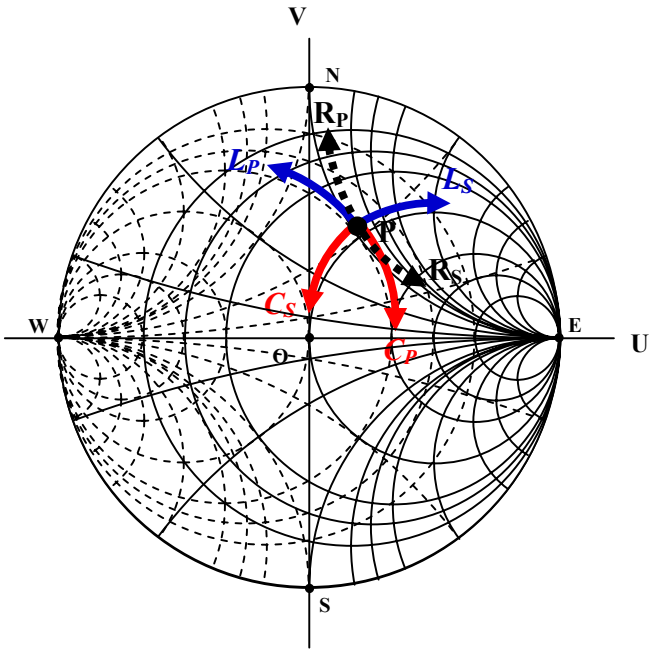


Figure 2.12 Pulled directions of impedance by the addition of L, C, or R on Smith Chart

It is usually impossible to match one port's impedance to a desired value, say, 50 Ω, by using only one part, either capacitor or inductor, with the exception of the special case when the original impedance to be matched is located on the circle with either r=1 or g=1. However, it is possible to use two parts to match one port's impedance to a desired value if the block with the matched port is operating for a narrow band system. The so-called "narrow band" means that the relative operating bandwidth is about less than 15%. Most of wireless communication systems are of narrow band. Before we discuss how the two parts are added to form a matching network, we would like to study the Smith Chart itself for a while.

2.3.2 Recognition of Regions in a Smith Chart

The topology of a matching network depends on the relative difference between the impedance to be matched to and the impedance to be matched from. If the destination impedance is 50 Ω, then the topology of the matching network depends only on the location of the impedance to be matched on the Smith Chart. For the matching purpose, the location of an impedance on a Smith Chart can be divided into 4 regions. Figure 2.13 shows the demarcation of these 4 regions on a Smith Chart:

- Region 1 is an area with low resistance or high conductance where
  $r < 1$, $x < |0.5|$,
  $g > 1$, $-\infty < b < +\infty$.
- Region 2 is an area with high resistance or low conductance where
  $r > 1$, $-\infty < x < +\infty$.
  $g < 1$, $b < |0.5|$.
- Region 3 is an area with low resistance and low conductance where
  $r < 1$, $x > 0$
  $g < 1$, $b < 0$.
- Region 4 is an area with low resistance and low conductance where
  $r < 1$, $x < 0$
  $g < 1$, $b > 0$.

The impedance in different regions is matched by means of different topologies. They will be discussed in a somehow detailed way in the following sections.

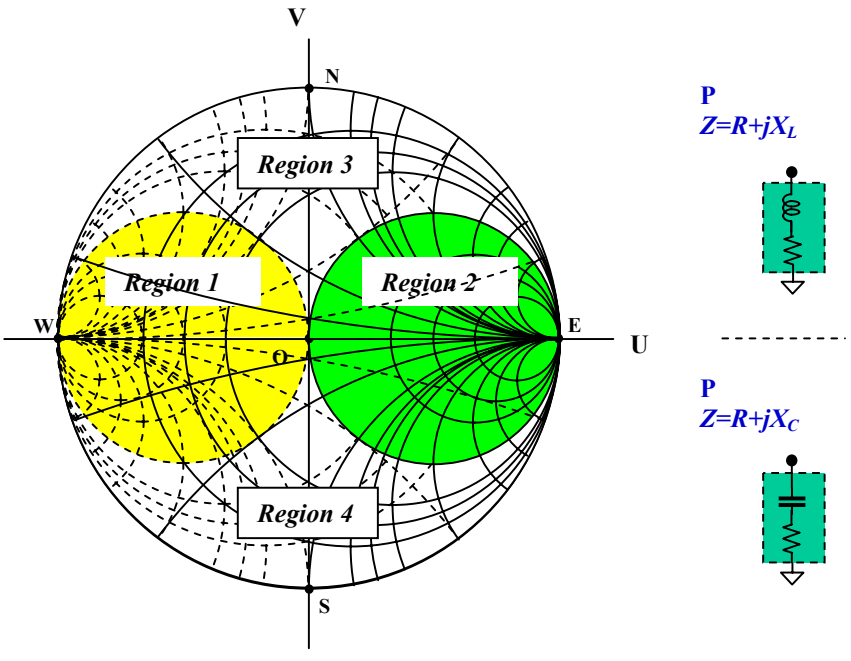


Figure 2.13 4 regions on a Smith Chart

2.3.3 Two Parts Matching Network

The impedance matching process is to pull an impedance to the desired reference impedance. The desired reference impedance is usually a pure resistor, say, 50 Ω, and is located at the center of the Smith Chart. Figure 2.14 shows that in regions 1 and 2 there are two ways to pull the original impedance, P_1, P_2, P_3, and P_4, to the center of the Smith Chart, O, by the addition of two types of passive parts, inductors and/or capacitors.

In region 1:

- One way as shown in Figure 2.14 is to pull P_1 to A by adding a capacitor C_S in series first, and then from A to O by adding of an inductor L_P in parallel secondly.
- Another way as shown in Figure 2.15 is to pull P_1 to B by adding an inductor L_S in series first, and then to pull B to O by adding a capacitor C_P in parallel secondly.
- One way as shown in Figure 2.14 is to pull P_2 to B by adding an inductor L_S in series first, and then from B to O by adding a capacitor C_P in parallel secondly.
- Another way as shown in Figure 2.15 is to pull P_2 to A by adding a capacitor C_S in series first, and then to pull A to O by adding an inductor L_P in parallel secondly.

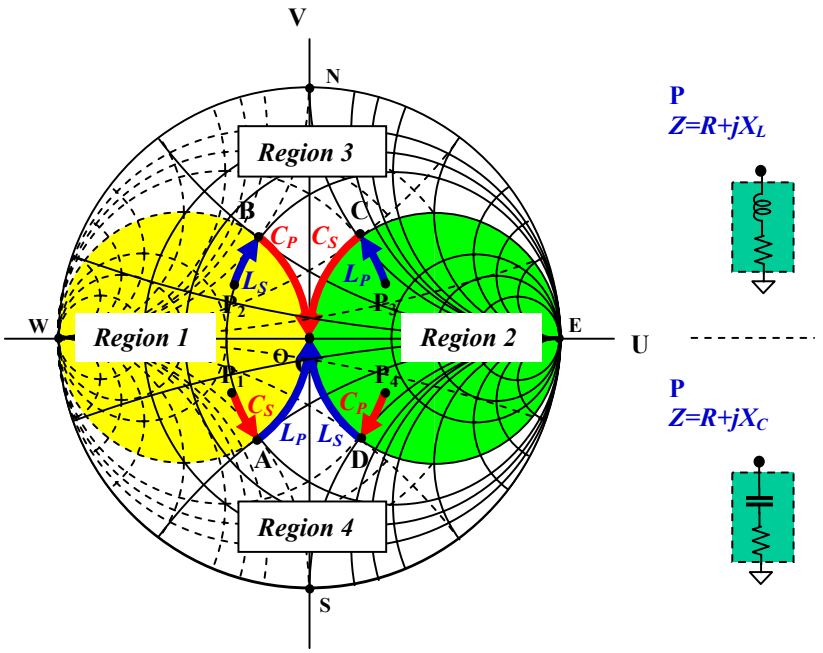


Figure 2.14 In regions 1 and 2, there are two ways to pull the original impedance P to the center of Smith Chart ,O, by addition of two passive parts

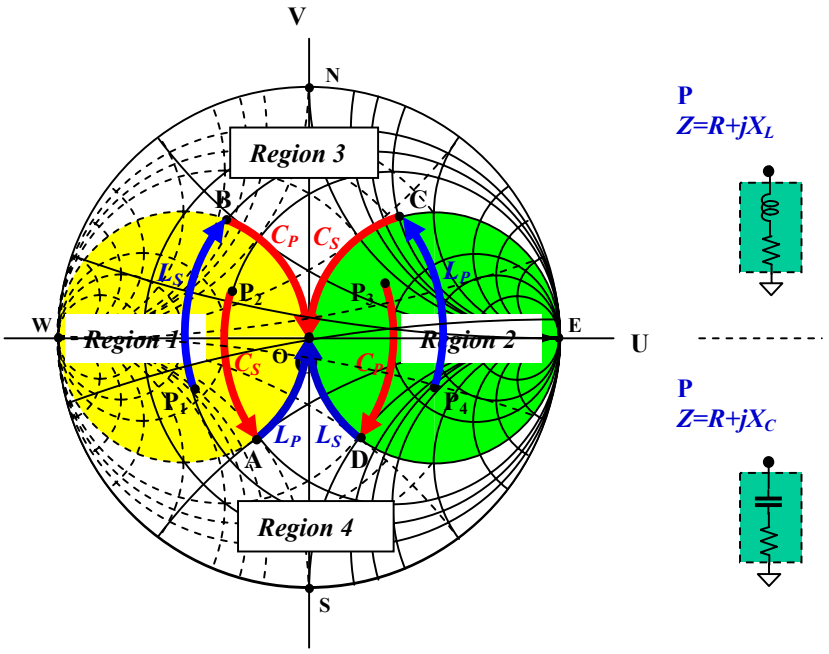


Figure 2.15 In regions 1 and 2, there are another two ways to pull the original impedance P to the center of Smith Chart ,O, by addition of two passive parts

It should be noted that in the matching ways described above there are two common features. They are:

1) The first matching part, either a capacitor or an inductor, is in series, and the second matching part, either a capacitor or an inductor, is in parallel.
2) The first matching part, either a capacitor or an inductor, is to pull the original impedance to the conductance circle, g=1, and the second matching part, either a capacitor or an inductor, is to pull the impedance from the conductance circle, g=1, to the reference impedance, or to the center of the Smith Chart.

In region 2:

- One way as shown in Figure 2.14 is to pull P_3 to C by adding an inductor L_P in parallel first, and then from C to O by adding a capacitor C_S in series secondly.
- Another way as shown in Figure 2.15 is to pull P_3 to D by adding a capacitor C_P in parallel first, and then to pull D to O by the addition an inductor L_S in series secondly.
- One way as shown in Figure 2.14 is to pull P_4 to D by adding a capacitor C_P in parallel first, and then from D to O by adding an inductor L_S in series secondly.
- Another way as shown in Figure 2.15 is to pull P_4 to C by adding an inductor L_P in parallel first, and then to pull C to O by the additing a capacitor C_S in series secondly.

Also, it should be noted that in these matching ways described above there are two common features. They are:

1) The first matching part, either a capacitor or an inductor, is in parallel, and the second matching part, either a capacitor or an inductor, is in series.
2) The first matching part, either a capacitor or an inductor, is to pull the original impedance to the resistance circle, r=1, and the second matching part, either a capacitor or an inductor, is to pull the impedance from the resistance circle, r=1, to the reference impedance, or to the center of the Smith Chart.

Figure 2.16 shows the topologies of matching networks described above. On the basis of the first feature, these topologies can be combined and are shown in Figure 2.17. The X_S and X_P are the reactance in series and in parallel respectively, and
when X is inductive,

$$X_S = X_{LS} = L_S\omega \quad , \tag{2.35}$$

$$X_P = X_{LP} = L_P\omega \quad , \tag{2.36}$$

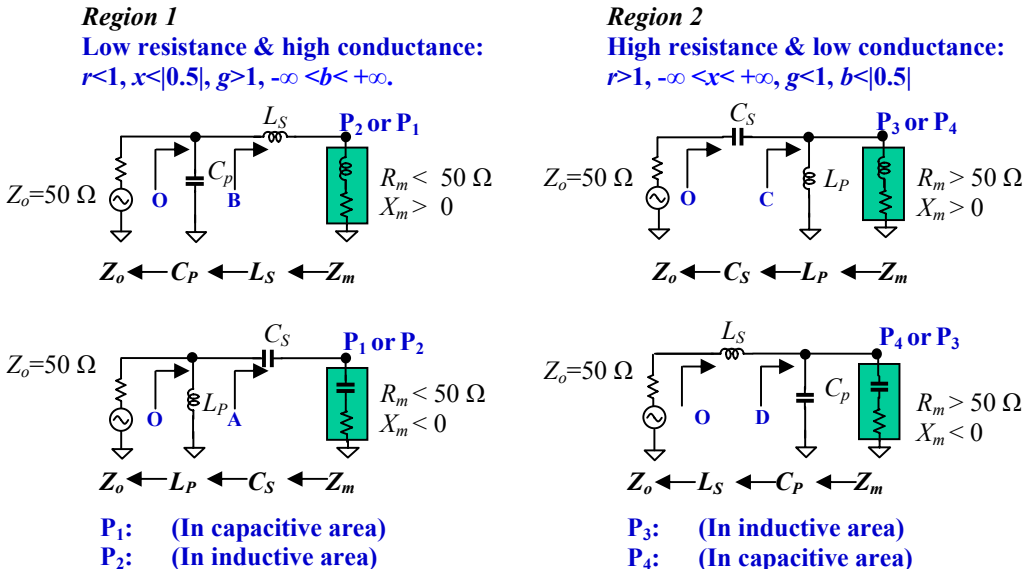


Figure 2.16 Impedance matching network for region 1 and 2 in Smith Chart

$Z = R + jX$,
$Y = G + jB = R/(R^2+X^2) - jX/(R^2+X^2)$

when X is capacitive,

$$X_S = X_{CS} = -\frac{1}{C_S\omega} \quad , \tag{2.37}$$

$$X_P = X_{CP} = -\frac{1}{C_P \omega} \quad . \tag{2.38}$$

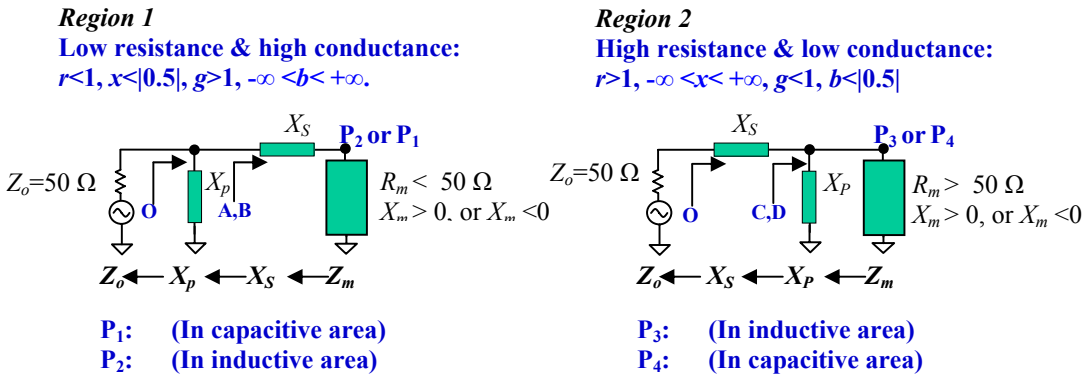


Figure 2.17 Impedance matching network for region 1 and 2 in Smith Chart

$Z = R + jX$,
$Y = G + jB = R/(R^2+X^2) - jX/(R^2+X^2)$

The impedance in region 1 is less than the reference impedance Z_o or 50 Ω, thus the impedance matching is called upward impedance matching. On the contrary, the impedance in region 2 is higher than the reference impedance Z_o or 50 Ω, thus the impedance matching is called downward impedance matching.

On the basis of the features mentioned above and based on Figure 2.17, we can calculate the values of the matching parts, one in series and another one in parallel. In order to do the calculation, let's introduce the normalized expressions of impedance and admittance:

$$Z = R + jX \ , \tag{2.39}$$

$$z = r + jx = \frac{Z}{Z_o} = \frac{R}{Z_o} + j\frac{X}{Z_o} \quad , \tag{2.40}$$

$$y = g + jb = \frac{1}{z} = \frac{1}{r + jx} = \frac{r - jx}{r^2 + x^2} \quad , \tag{2.41}$$

$$g = \frac{r}{r^2 + x^2} \quad , \tag{2.42}$$

$$b = -\frac{x}{r^2 + x^2} \quad . \tag{2.43}$$

In Figure 1.6 and 1.7, Z_m denotes the original impedance to be matched. It consists of real and imaginary part R_m and X_m, that is,

$$Z_m = R_m + jX_m \quad . \tag{2.44}$$

Region 1 is an upward matching area, where

$$R_m < Z_o = 50 \ \ \Omega \, , \tag{2.45}$$

The real and imaginary part of the impedance at point A or B, $r_{A,B}$ and $x_{A,B}$, after adding the first matching part X_S in series, are

$$r_{A,B} = r_m \quad , \tag{2.46}$$

and

$$x_{A,B} = x_S + x_m \ , \tag{2.47}$$

respectively. The first matching part, X_S, in series, is to move the original impedance, Z_m, to the circle with $g = g_{A,B} = 1$. Consequently,

$$g = g_{A,B} = \frac{r_{A,B}}{{r_{A,B}}^2 + {x_{A,B}}^2} = \frac{r_m}{{r_m}^2 + \left(x_S + x_m\right)^2} = 1 \quad . \tag{2.48}$$

From equation (2.48), we have

$${r_m}^2 + \left(x_S + x_m\right)^2 = r_m \ , \tag{2.49}$$

$$\left(\frac{R_m}{Z_o}\right)^2 + \left(\frac{X_S + X_M}{Z_o}\right)^2 = \frac{R_M}{Z_o} \quad , \tag{2.50}$$

$$X_S + X_m = \pm\sqrt{R_m\left(Z_o - R_m\right)} \quad , \tag{2.51}$$

$$X_S = \pm\sqrt{R_m\left(Z_o - R_m\right)} - X_m \quad . \tag{2.52}$$

The positive sign in equation (2.52) is for the selection of an inductive X_S while the negative sign is for the selection of a capacitive X_S.

The second matching part, X_P, in parallel, is to pull $b_{A,B}$ at point A or B to zero, that is,

$$\frac{1}{x_P} + \left(-b_{A,B}\right) = 0 \quad , \tag{2.53}$$

From expressions (2.43), (2.46), (2.47) and (2.53) we have

$$\frac{1}{x_P} = b_{A,B} = -\frac{x_S + x_m}{r_m{}^2 + (x_S + x_m)^2} \quad . \tag{2.54}$$

From (2.52) and (2.54), we have

$$\frac{1}{X_P} = \frac{X_S + X_m}{R_m{}^2 + (X_S + X_m)^2} = \frac{\pm\sqrt{R_m(Z_o - R_m)}}{R_m{}^2 + R_m(Z_o - R_m)} \quad , \tag{2.55}$$

$$\frac{1}{X_P} = \frac{\pm\sqrt{R_m(Z_o - R_m)}}{R_m Z_o} \quad , \tag{2.56}$$

The positive sign in equation (2.56) is for the selection of inductive X_P while the negative sign is for the selection of capacitive X_P.

In terms of equations (2.52) and (2.56), the values of the two matching parts, X_S and X_P, can be calculated for region 1. Also, a simple relationship between X_S and X_P can be found from equation (2.52) and (2.56):

$$X_P(X_S + X_m) = R_m Z_o \quad . \tag{2.57}$$

As an example, let's apply equations (2.52) and (2.56) to the impedance P_2 as shown in Figure 2.14, which corresponding to the topology as shown in the left upper corner in Figure 2.16. The first matching part, L_S, and the second matching part, C_p, for P_2, is

$$L_S = \frac{\sqrt{R_m(Z_o - R_m)} - X_m}{\omega} \quad , \tag{2.58}$$

and

$$C_P = \frac{\sqrt{R_m(Z_o - R_m)}}{R_m Z_o \omega} \quad , \tag{2.59}$$

respectively.

As to the impedance P_1 as shown in Figure 2.14, which corresponds to the topology as shown in the left bottom corner in Figure 2.16. The first matching part C_S and the second matching part L_p for P_1, is

$$C_S = \frac{C_m}{C_m \omega \sqrt{R_m(Z_o - R_m)} - 1} \quad , \tag{2.60}$$

and

$$L_P = \frac{R_m Z_o}{\omega\sqrt{R_m(Z_o - R_m)}} \quad , \tag{2.61}$$

respectively.

Region 2 is a downward matching area, where

$$R_m > Z_o = 50\ \Omega\ , \tag{2.62}$$

The impedance at point C or D, $Z_{C,D}$, after adding the first matching part X_P in parallel, is

$$Z_{C,D} = \frac{(R_m + jX_m)jX_P}{(R_m + jX_m) + jX_P} = \frac{(-X_mX_P + jX_PR_m)[R_m - j(X_m + X_P)]}{R_m{}^2 + (X_m + X_P)^2}\ , \tag{2.63}$$

$$R_{C,D} = \frac{R_mX_P{}^2}{R_m{}^2 + (X_m + X_P)^2}\ , \tag{2.64}$$

$$X_{C,D} = \frac{R_m{}^2 + X_m(X_m + X_P)}{R_m{}^2 + (X_m + X_P)^2}X_P\ . \tag{2.65}$$

The first matching part X_P in parallel, is to get $r = r_{C,D} = 1$, or $R_{C,D} = Z_o$. Consequently, from (2.64), we have

$$R_{C,D} = \frac{R_mX_P{}^2}{R_m{}^2 + (X_m + X_P)^2} = Z_o\ , \tag{2.66}$$

$$(R_m - Z_o)X_P{}^2 - 2X_mZ_oX_P - (R_m{}^2 + X_m{}^2)Z_o = 0\ , \tag{2.67}$$

$$X_P = \frac{X_mZ_o \pm \sqrt{R_mZ_o(R_m{}^2 + X_m{}^2 + R_mZ_o)}}{R_m - Z_o}\ , \tag{2.68}$$

The positive sign in equation (2.68) is for the selection of inductive X_P while the negative sign is for the selection of capacitive X_P.

The second matching part X_S in series is to pull $X_{A,B}$ at point C or D to zero, that is,

$$X_S = -X_{C,D} = -\frac{R_m{}^2 + X_m(X_m + X_P)}{R_m{}^2 + (X_m + X_P)^2}X_P = \left[\frac{(X_m + X_P)Z_o}{R_mX_P} - 1\right]X_P\ , \tag{2.69}$$

$$X_S = \frac{Z_o}{R_m}X_m + \frac{Z_o - R_m}{R_m}X_P\ , \tag{2.70}$$

$$X_S = \mp \frac{\sqrt{R_m Z_o \left({R_m}^2 + {X_m}^2 - R_m Z_o \right)}}{R_m} \quad . \qquad (2.71)$$

The positive sign in equation (2.71) is for the selection of inductive X_S while the negative sign is for the selection of capacitive X_S.

In terms of equations, (2.68) and (2.71), the values of two matching parts, X_P and X_S, can be calculated for the region 2. As an exercise, readers are able to derive 4 equations similar to the equations (2.58) to (2.61), by which the values of L_P, C_S, or C_P, L_S for matching of P_3 or P_4 in region 2 as shown in Figure 2.16 can be calculated. Also, a simple relationship between X_P and X_S can be found from equation (2.68) and (2.71):

$$\left[X_m Z_o - X_P \left(R_m - Z_o \right) \right] X_S = Z_o \left({R_m}^2 + {X_m}^2 - R_m Z_o \right) \quad . \qquad (2.72)$$

Now let's turn to the regions 3 and 4.

Figure 2.18 shows that in regions 3 and 4 there are two ways to pull the original impedance P to the center *O* of the Smith Chart by the addition of two passive parts.

Let's denote the desired impedance to be matched, P_5 and P_6, for region 3 and region 4 respectively.

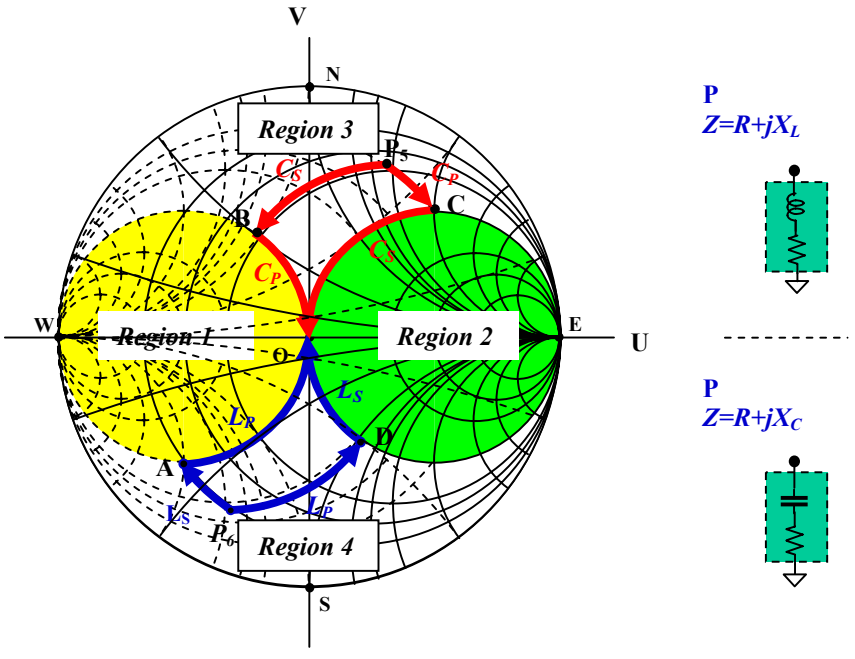


Figure 2.18 In region 3 and 4, there are two ways to pull the original impedance P to the center *O* of Smith Chart ,O, by addition of two passive parts.

In region 3:

- One way as shown in Figure 2.18 is to pull P_5 to B by adding a capacitor C_S in series first, and then from B to O by adding a capacitor C_P in parallel.
- Another way as shown in Figure 2.18 is to pull P_5 to C by adding a capacitor C_P in parallel first, and then to pull C to O by adding a capacitor C_P in series.

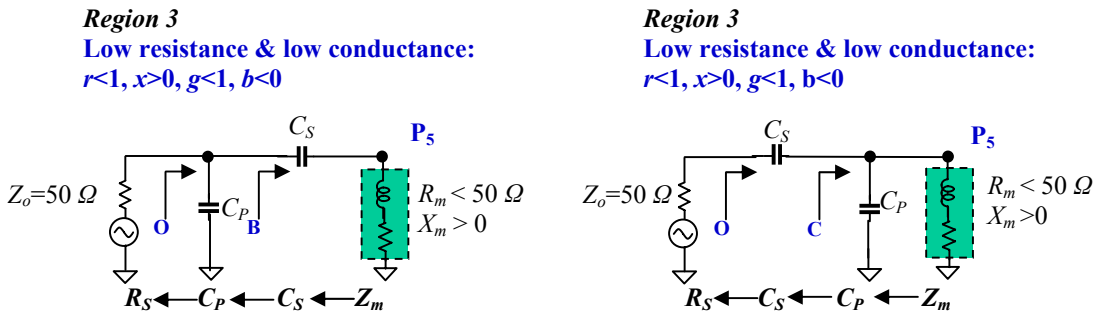
It should be noted that the matching ways described above have two common features. They are:

1) Both of the two matching parts are capacitive.
2) The first matching part, either in series or in parallel, is to pull the original impedance to the conductance circle, g=1, or to the resistance circle, r=1. And, the second matching part, either in parallel or in series, is to pull the impedance from the conductance circle, g=1, or from the resistance circle, r=1, to the reference impedance, or the center of the Smith Chart.

One of the two ways must be decided by the designer. The criteria depends on the value of the capacitors:

- The value must not be too low. It must be much higher than the spray value in the layout;
- The value must not be too high. A capacitor with too high value would cost too much die area in the RFIC design, or bring about the current leakage in the discrete RF design.

Figure 2.19 sketches the topology of the impedance matching for the original impedance located in region 3.



(a) First matching part is in series
Second matching part is in parallel

(b) First matching part is in parallel
Second matching part is in series

P_5 (In inductive area)

Figure 2.19 Impedance matching network for the original impedance located in region 3 of the Smith Chart

$Z = R + jX$,

$Y = G + jB = R/(R^2+X^2) - jX/(R^2+X^2)$

In region 4:

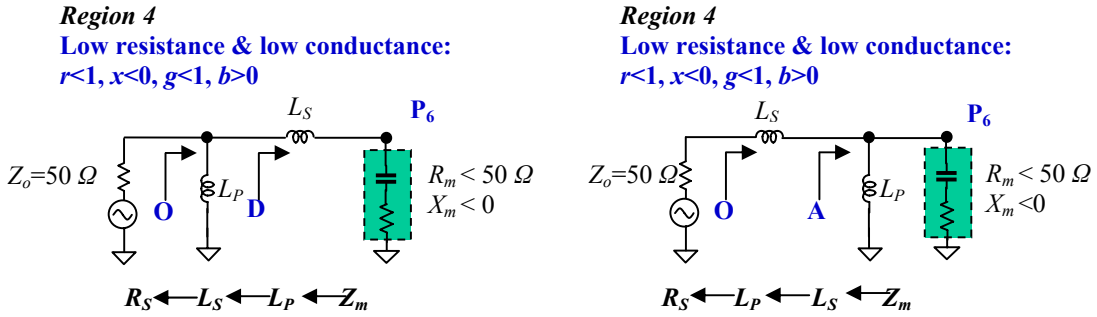
- One way as shown in Figure 2.20 is to pull P_6 to D by adding an inductor L_P in parallel first, and then from D to O by adding an inductor L_P in series.
- Another way as shown in Figure 2.20 is to pull P_6 to A by additing an inductor L_S in series first, and then to pull A to O by adding an inductor L_P in parallel.

It should be noted that the matching ways described above have two common features. They are:

1) Both of the two matching parts are inductive.
2) The first matching part, either in series or in parallel, is to pull the original impedance to the conductance circle, g=1, or to the resistance circle, r=1. And, the second matching part, either in parallel or in series, is to pull the impedance from the conductance circle, g=1, or from the resistance circle, r=1, to the reference impedance, or the center of Smith Chart.

- One of the two ways must be decided by the designer. The criteria depends on the value of the capacitors:
- The value must not be too low. It must be higher than the minimum of inductor which is available in the component cabinet;
- The value must not be too high. An inductor with too high value would cost too much die area in the RFIC design.

Figure 2.20 sketches the topology of the impedance matching for the original impedance located in region 4.



(a) First matching part is in series
Second matching part is in parallel

(b) First matching part is in parallel
Second matching part is in series

P_6 (In capacitive area)

Figure 2.20 Impedance matching network for the original impedance located in region 4 of the Smith Chart

$Z = R + jX$,

$Y = G + jB = R/(R^2+X^2) - jX/(R^2+X^2)$

As a matter of fact, expressions (2.52) and (2.56) are available to all the cases when the first matching part is in series and the second matching part is in parallel. It can therefore be used to calculate the two matching parts in the corresponding cases in region 3 and

region 4, as shown at the left portion (a) in Figure 2.19 and Figure 2.20. On the other hand, the expressions (2.68) and (2.71) are available to all the cases when the first matching part is in parallel and the second matching part is in series. It can therefore be used to calculate the two matching parts in the corresponding cases in region 3 and region 4, as shown at the right portion (b) in Figure 2.19 and Figure 2.20.

It is a remarkable feature if a matching network consists of only capacitors. This is the case if the original impedance is located in region 3. It is a good deal to the design engineer because a capacitor is much easier to deal with than an inductor from the viewpoint of the cost and the quantity factor. On the contrary, it would be a significant disadvantage to build a matching network for the original impedance located in region 4, which consists only of inductors. Unfortunately, the input impedance of a CMOS device is usually located in the region 4.

So far, we have discussed two parts matching network in all of 4 regions of a Smith Chart. Let's summarize their key points as follows:

- The 1st component is to move the impedance to the circle with either g=1 or r=1.
- The 2nd component is to pull the intermediate impedance from the circle with either g=1 or r=1 to the center of the Smith Chart, that is, to the point of the reference resistance and zero reactance.
- Matching network in region 1 is an upward impedance transformer.
  * It can transfer the impedance from low to high (50 Ω here).
  * Usually it is the case in the design of power amplifier.
- Matching network in region 2 is a downward impedance transformer.
  * It can transfer the impedance from high to low (50 Ω here).
  * Usually it is the case in the design of an LNA, mixer etc. in a CMOS process.
- Matching network in region 3 and 4 is a simple matching unit with only one type of parts.
  * In region 3 it can be matched only by capacitors,
    One is in series and another in parallel.
    The order of these two capacitors doesn't matter.
  * In region 4 it can be matched only by inductors,
    One is in series and another in parallel.
    The order of these two inductors doesn't matter.

2.3.4 Two Parts Upward and Downward Impedance Transformer

In the discussion above, the original impedance Z_m consists of real and imaginary parts, R_m and X_m. Now let's introduce a special case in which both of the original impedance Z_m and the reference or final impedance Z_o are pure resistive parts, that is,

$$Z_m = R_m \quad , \qquad X_m = 0 \quad , \tag{2.73}$$

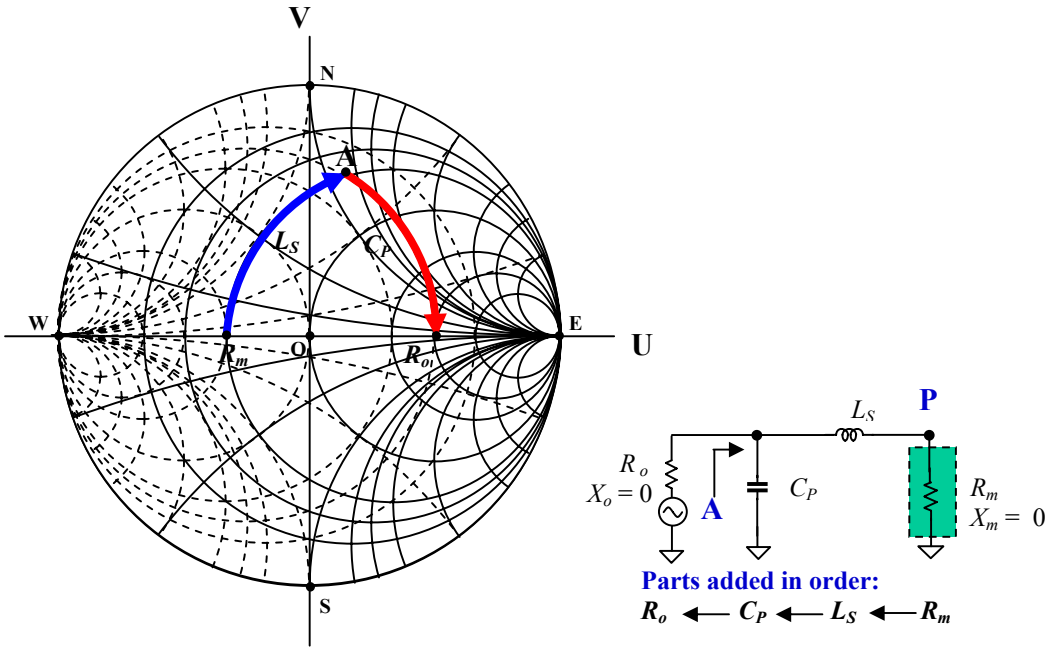


Figure 2.21 Upward impedance transformer with C_P and L_S.

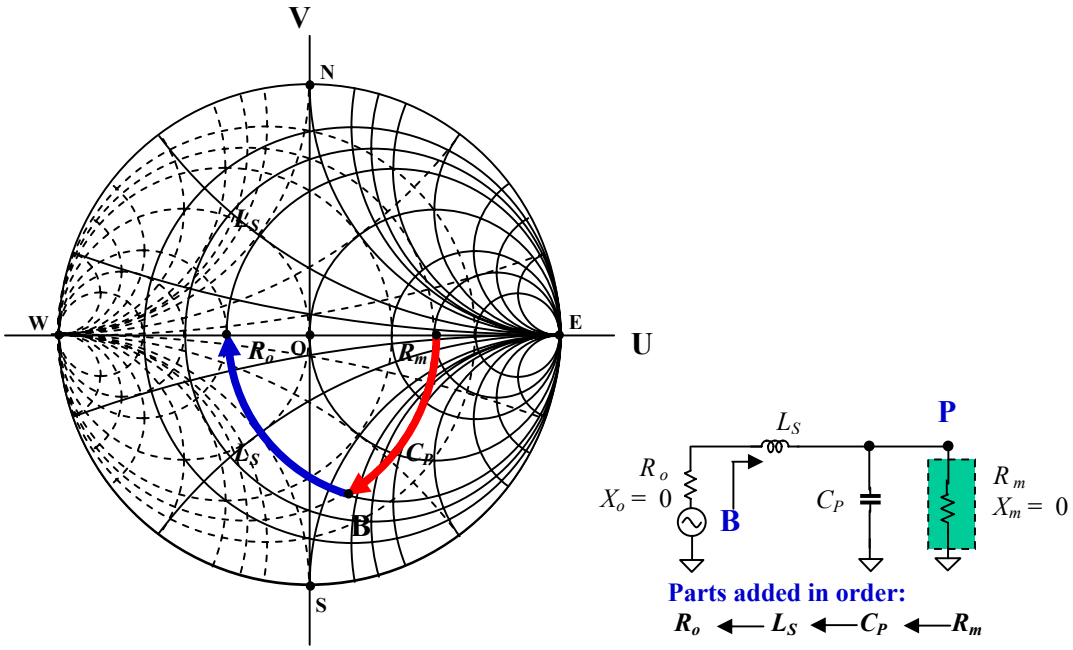


Figure 2.22 Downward impedance transformer with C_P and L_S.

$$Z_o = R_o \ , \qquad X_o = 0 \ , \tag{2.74}$$

where it is not necessary for R_o to be 50 Ω. In such cases the matching network, in fact, is a transformer with the impedance transformation between R_m and R_o.

Figure 2.21 shows an upward transformer with C_P and L_S, in which

$$R_m < R_o \ , \tag{2.75}$$

while Figure 2.22 shows a downward transformer with C_P and L_S, in which

$$R_m > R_o \ . \tag{2.76}$$

They are, in fact, imaginary from each other. In other words, both of them become the same if we exchange the position of the load and the source, R_m and R_o, from the upward to downward transformer. Consequently, their analytical expressions are the same.

Let's analyse the downward impedance transformer as shown in Figure 2.22. The resistor R_o in series with L_S and the resistor R_m in parallel with C_P are impedance-matched if their Q values, expressed by (2.26) and (2.30) respectively, are equal, such as

$$Q = \omega_o C_P R_m = \frac{\omega_o L_S}{R_o}, \tag{2.77}$$

where

$$\omega_o^{\,2} = \frac{1}{C_P L_S} \ , \tag{2.78}$$

and

$$R_m = R_o\left(1+Q^2\right) = R_o\left[1+\frac{1}{\left(\omega_o R_o C_P\right)^2}\right] = R_o + \frac{1}{R_o}\frac{L_S}{C_P}, \tag{2.79}$$

then,

$$R_o\left(R_m - R_o\right) = \frac{L_S}{C_P} \ . \tag{2.80}$$

where Q is called "quality factor" and Z_o is the characteristic impedance of the matching network.

From equations (2.78) and (2.80), we have

$$C_P = \frac{1}{\omega_o \sqrt{R_o\left(R_m - R_o\right)}} \quad , \tag{2.81}$$

$$L_S = \frac{\sqrt{R_o\left(R_m - R_o\right)}}{\omega_o} \quad , \tag{2.82}$$

In these two equations, R_o, R_m, and ω_o, are given parameters. The values of C_P and L_S can be calculated from equations (2.81) and (2.82). They are real numbers because in a downward transformer, $R_m > R_o$, as shown in the expression (2.76).

Now let's shift to the upward impedance transformer with C_P and L_S as shown in Figure 2.21. As mentioned above, the expressions for C_P and L_S can be obtained by simply exchanging the positions of the load and the source, R_m and R_o, in the equations (2.81) and (2.82), that is

$$C_P = \frac{1}{\omega_o \sqrt{R_m (R_o - R_m)}} , \qquad (2.83)$$

$$L_S = \frac{\sqrt{R_m (R_o - R_m)}}{\omega_o} . \qquad (2.84)$$

In these two equations, R_o, R_m, and ω_o, are given parameters. The values of C_P and L_S can be calculated from equations (2.83) and (2.84). They are real numbers because in an upward transformer, $R_m < R_o$, as shown in the expression (2.75).

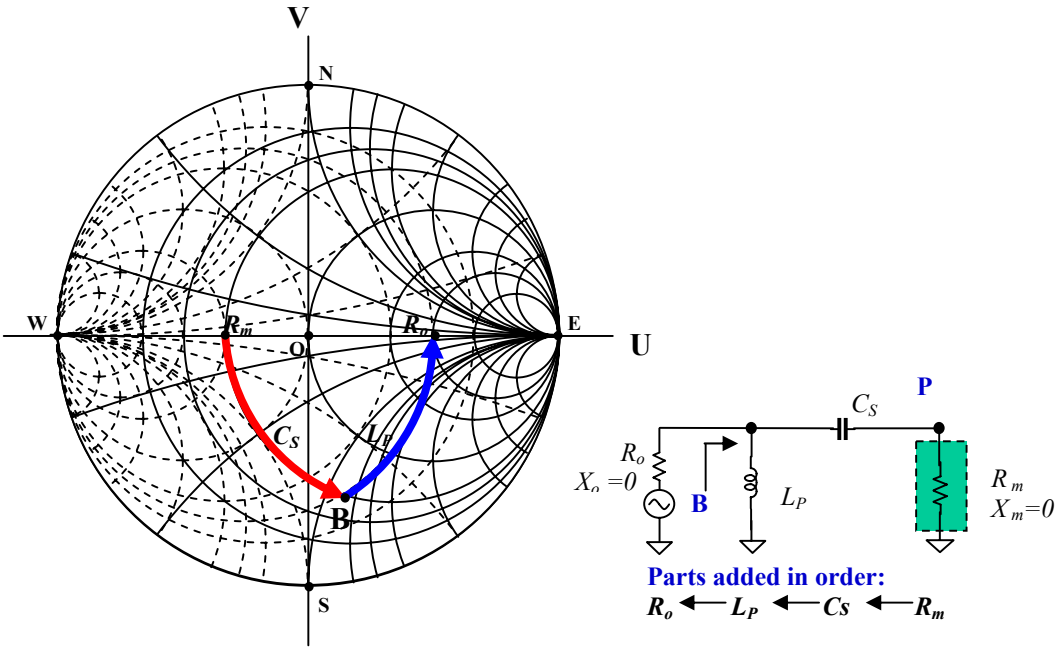


Figure 2.23 Upward impedance transformer with C_S and L_P.

As a matter of fact, the upward and downward impedance transformer as shown in Figure 2.21 or 2.22 is a basic segment of a low pass filter. We can have another form of the matching network between R_o and R_m. Figure 2.23 and 2.24 show an alternative upward and downward impedance transformer, or, in other words, a matching network with a basic segment of a high pass filter.

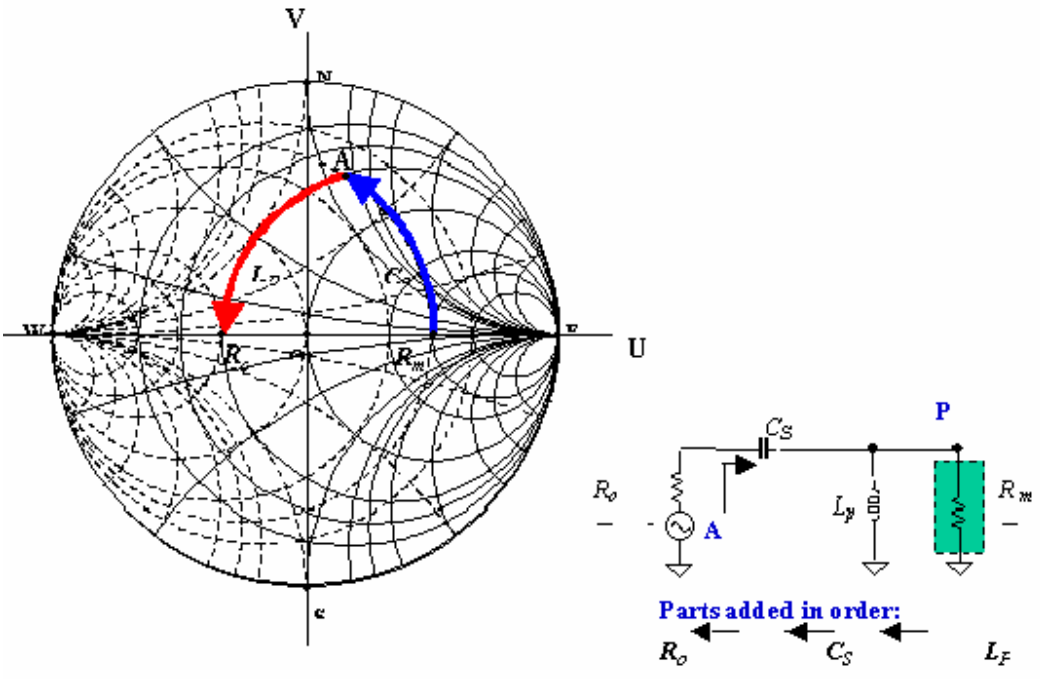


Figure 2.24 Downward impedance transformer with C_S and L_P.

Again, as shown in Figures 2.23 and 2.24, these two types of upward and downward impedance transformer are imaginary from each other. That is, both of them become the same if we exchange the position of the load and the source, R_o and R_m, from the upward to downward impedance transformer. The derivation of the two analytical expressions for C_S and L_P, which is similar to expressions, (2.81), (2.82), (2.83), and (2.84), will be remained to readers as an exercise work.

2.3.5 Three Parts Matching Network and Impedance Transformer

2.3.5.1 Topology Limitation of Two Parts Matching Network

A matching network constructed by only two parts, one capacitor and one inductor, is quite a popular selection. Figure 2.25 shows 4 different topologies, or, in other words, 4

different combinations of one capacitor and one inductor in series or in parallel by means of two Smith Charts.

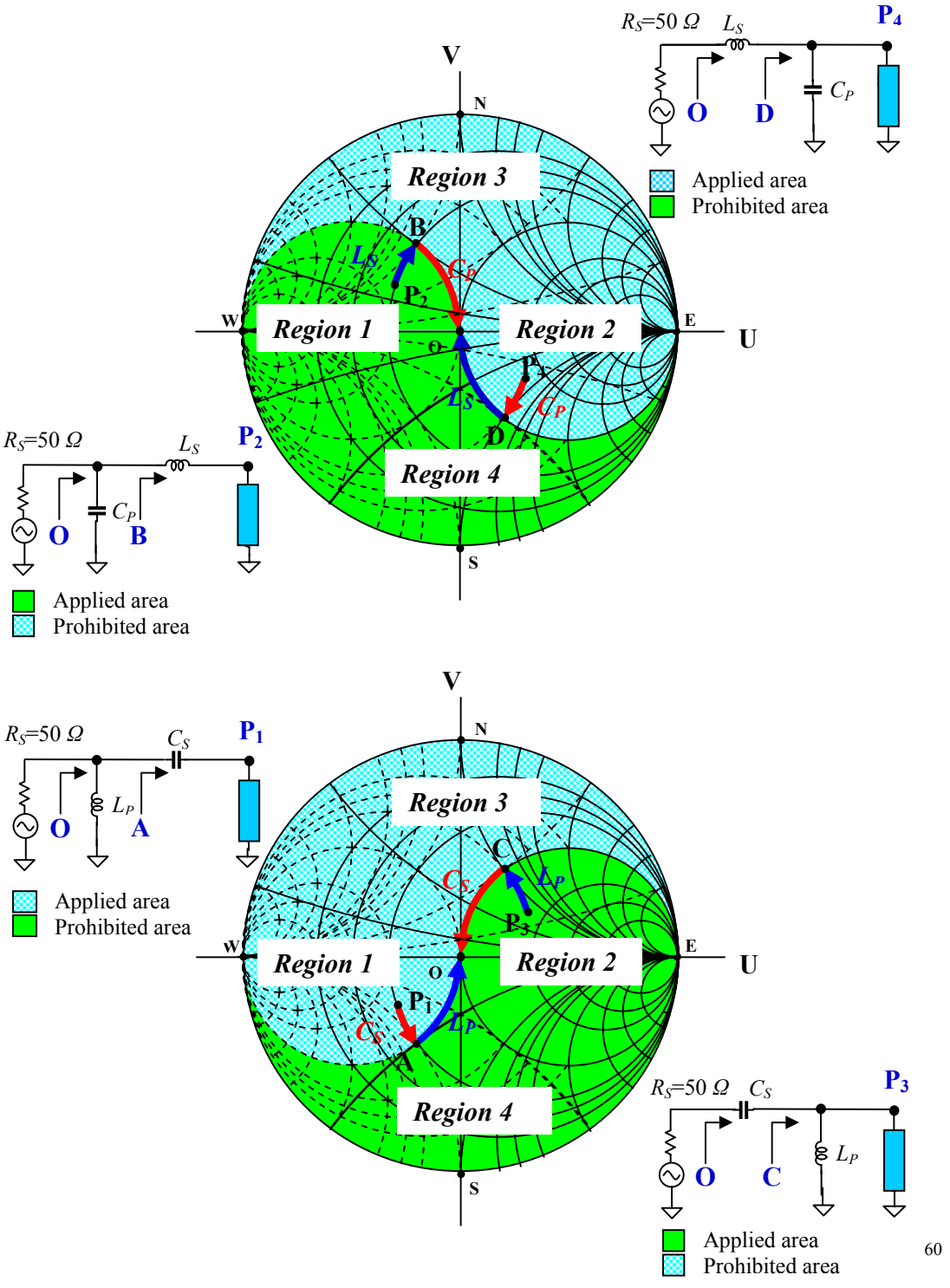


Figure 2.25 Prohibitive area for different topology of a matching network with two parts

In the top Smith Chart of Figure 2.25, the topology for P_2 can be applied to both regions 1 and 4 but not to regions 2 and 3. In other words, this matching topology is available only for regions 1 and 4. Regions 2 and 3 are the prohibited areas of the matching topology for P_2. Any impedance in region 2 or 3 cannot be matched to the center *O* of the Smith Chart through the topology for P_2. On the contrary, the topology for P_4 can be applied to both regions 2 and 3 but not to regions 1 and 4. Any impedance in regions 1 or 4 cannot be matched to the center *O* of the Smith Chart through the topology for P_4.

Similarly, in the bottom Smith Chart of Figure 2.25, the topology for P_1 can be applied to both regions 1 and 3 but not to regions 2 and 4. In other words, this matching topology is available only for regions 1 and 3. Regions 2 and 4 are the prohibited areas of the matching topology for P_1. Any impedance in regions 2 or 4 cannot be matched to the center *O* of the Smith Chart through the topology for P_1. On the contrary, the topology for P_3 can be applied to both regions 2 and 4 but not to regions 1 and 3. Regions 1 and 3 are the prohibited areas of the matching topology for P_3. Any impedance in region 1 or 3 cannot be matched to the center *O* of the Smith Chart through the topology for P_3.

From the viewpoint of simplicity and cost, a matching network should be implemented by as few parts as possible. In practical engineer designing, there is a very slim chance to build a matching network containing only one part whereas it is quite conventional to implement a matching network by two parts although there is a restriction as shown in Figure 2.25. Such a restriction, however, could be avoided as long as the designer selects a correct topology.

On the other hand, a matching network constructed by three parts, a combination of capacitors and inductors, can bring such a restriction down significantly. Of course, the price is to pay for an extra part.

All the topologies of a matching network constructed by 3 parts can be categorized into *Π* and *T* types. Let's briefly discuss them one by one.

2.3.5.2 Π Type Matching Network

A Π type matching network contains 3 matching parts. There are 8 possible alternative topologies:

a) C_{P1}-C_S-C_{P2},
b) C_{P1}-C_S-L_{P2},

c) C_{P1}-L_S-C_{P2},
d) C_{P1}-L_S-L_{P2},
e) L_{P1}-C_S-C_{P2},
f) L_{P1}-C_S-L_{P2},
g) L_{P1}-L_S-C_{P2},
h) L_{P1}-L_S-L_{P2}.

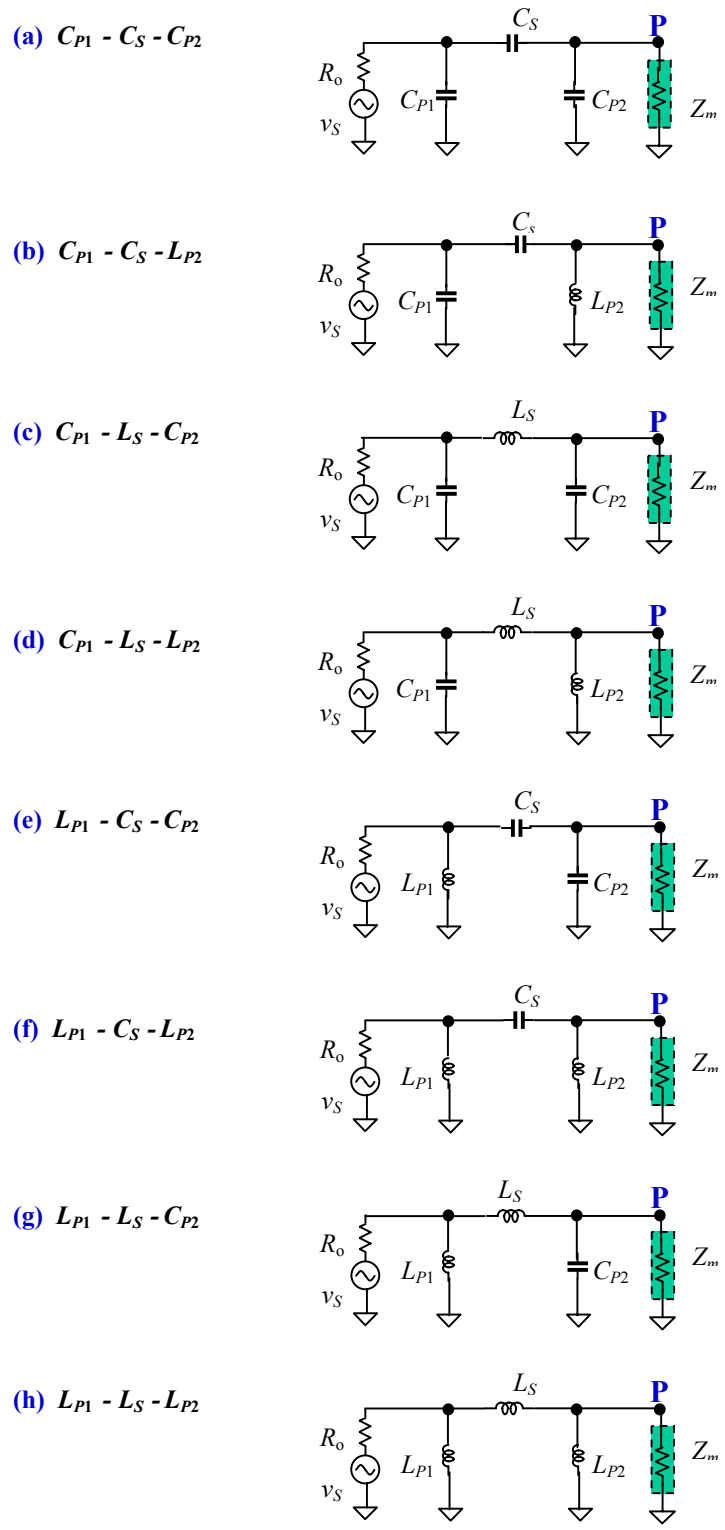


Figure 2.26 Topologies of a *Π* type matching network

Table 2.1 Advantages and disadvantages of different topologies in a *Π* type matching network

Topology	Inductor number	Advantages	Disadvantages
(a) C_{p1} **-** C_S **-** C_{p2}	0	Lowest cost	Region 3 only
(b) C_{p1} **-** C_S **-** L_{p2}	1	Low cost	DC short-circuited Region 2,3,4
(c) C_{p1} **-** L_S **-** C_{p2}	1	Low cost	DC un-blocked Region 1,4
(d) C_{p1} **-** L_S **-** L_{p2}	2		Medium cost DC un-blocked DC short-circuited
(e) L_{p1} **-** C_S **-** C_{p2}	1	Low cost	Region 1, 3
(f) L_{p1} **-** C_S **-** L_{p2}	2		Medium cost DC short-circuited
(g) L_{p1} **-** L_S **-** C_{p2}	2		Medium cost DC un-blocked
(h) L_{p1} **-** L_S **-** L_{p2}	3		Highest cost DC short-circuited DC un-blocked

Figure 2.26 shows 8 possible topologies of a *Π* type matching network and Table 2.1 list their advantages and disadvantages.

It should be noted that the source, with an impedance of R_o=50 *Ω*, and the original impedance Z_m is matched by inserting a matching network between them. In practical circuit designs, the original impedance Z_m is the input impedance of the next stage. Therefore, in the selection of a matching network topology, its DC blocking and DC short-circuiting resistance capabilities are the important factors. Another important factor, of course, is the cost.

Topology a), C_{P1}-C_S-C_{P2}, contains no inductor but 3 capacitors. The advantages of this topology are:

1) There is no DC short-circuited and DC blocking problem;

2) It costs the least out of all the topologies because a capacitor is usually much cheaper than an inductor;
3) The die area of an inductor is usually much larger than that of a capacitor in an RFIC design, so it has the smallest die area;
4) In an RFIC design the Q value of a capacitor is much higher than that of an inductor.

It, of course, is the most welcome topology. However, it should be noted that this topology is only available for the original impedance which is located in region 3 of Smith Chart.

Another 3 topologies of the above eight contain only one inductor. They are: b) C_{P1}-C_S-L_{P2}, c) C_{P1}-L_S-C_{P2}, and e) L_{P1}-C_S-C_{P2}. Unfortunately, a DC short-circuited problem exists in topology b) C_{P1}-C_S-L_{P2}, because the input of the next stage will be directly DC short-circuited to the ground by the inductor L_{P2}. Also, a DC blocking problem exists in the topology c) C_{P1}-L_S-C_{P2}, because the input of the next stage will be directly DC short-circuited to the source by the inductor L_S. In contrast to these two topologies, topology e) L_{P1}-C_S-C_{P2} is a better one, in which there is neither DC blocking nor DC short-circuited problem to next stage. In addition, its capacitor C_{P2} would alleviate tolerance due to the input spray capacitance of next stage. This topology is sometimes called as the "tapped" capacitor matching network.

The last 4 topologies contain two or three inductors. They are topologies d), f), g), and h). Usually they are not considered and adapted in the circuit design because they are too expensive.

In a practical design, an appropriate topology must be selected among those topologies that contain one or no inductors. The candidates are a) C_{P1}-C_S-C_{P2}, b) C_{P1}-C_S-L_{P2}, c) C_{P1}-L_S-C_{P2}, and e) L_{P1}-C_S-C_{P2}. The selection, of course, depends on the location of the original impedance Z_m on the Smith Chart because there is a regional prohibition for all these topologies. Figure 2.27 to 2.30 shows the available and prohibited areas on the Smith Chat for these 4 topologies. Sometimes an additional DC-blocking capacitor should be added whenever the problem of DC short-circuited or un-blocked DC occurs in the selected topology.

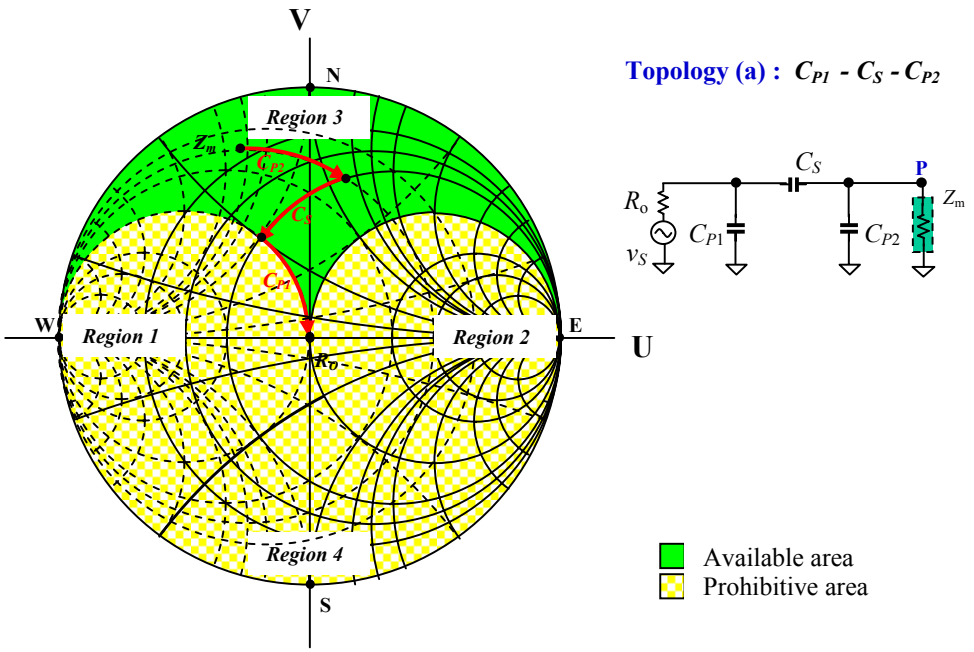


Figure 2.27 Available and prohibitive area of the topology: C_{P1}-C_S-C_{P2}.

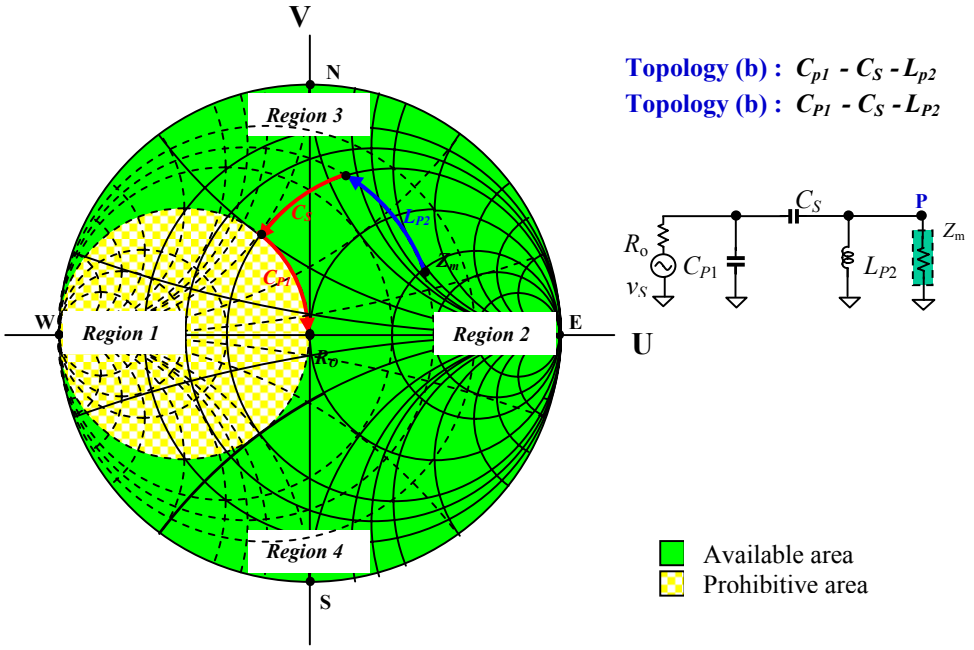


Figure 2.28 Available and prohibitive area of the topology: C_{P1}-C_S-L_{P2}.

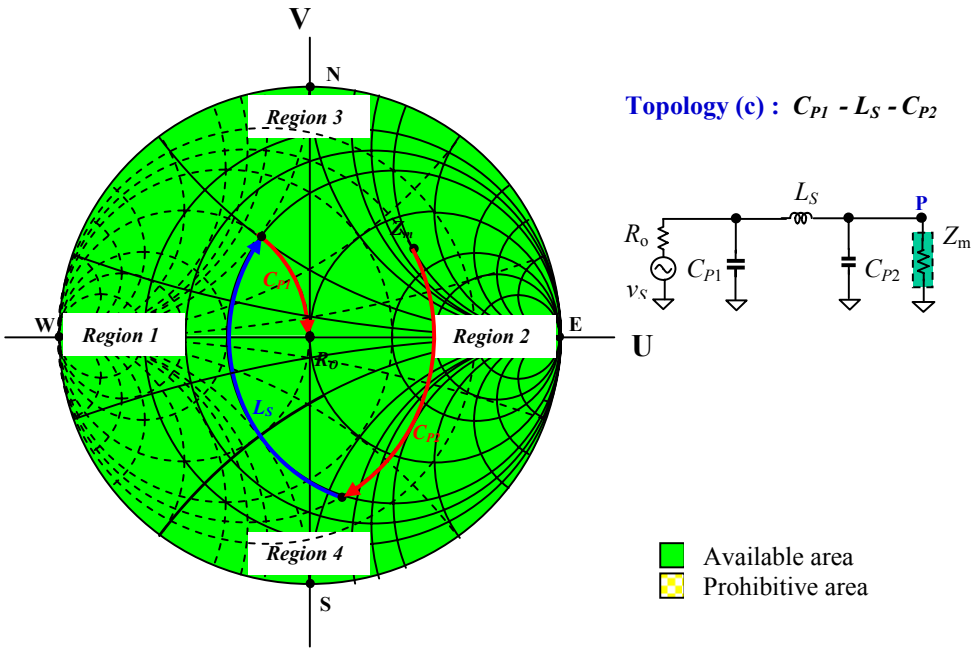


Figure 2.29 Available and prohibitive area of the topology: C_{P1}-L_S-C_{P2}.

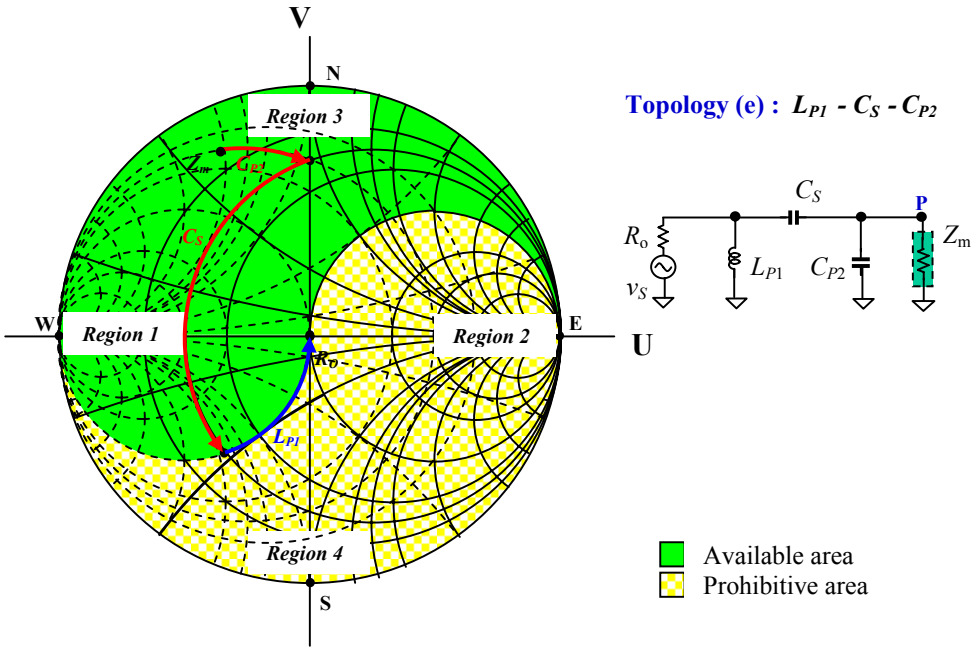


Figure 2.30 Available and prohibitive area of the topology: L_{P1}-C_S-C_{P2}.

From Figure 2.27 to 2.30 it can be seen that the topology a) C_{P1}-C_S-C_{P2} can be applied only when the original impedance Z_m is located in region 3. Topology b) C_{P1}-C_S-L_{P2} can be applied when the original impedance Z_m is located in region 2, 3, 4 but not in region 1. Topology c) C_{P1}-L_S-L_{P2} can be applied for any value of the original impedance Z_m, so that it is easiest to be used for a designer. Topology e) L_{P1}-C_S-C_{P2} can be applied when the original impedance Z_m is located in region 1 and 3. Usually the input impedance of a MOSFET device is located in region 2 or 4 so that its input impedance matching network is usually implemented by topology (c) or (b), while the input impedance of a bipolar is usually located in region 1 or 3 so that its input impedance matching network is usually implemented by topology (c) or (e).

It must point out that in a *Π* matching network containing 3 parts, there are infinitively possible moving paths of the impedance for each topology on the Smith Chat. For example, Figure 2.31 shows three possible moving paths of impedance on the Smith Chart, which corresponds to three possible values of C_{P1}, L_S, *and* C_{P2} for the topology C_{P1}- L_S - C_{P2}. It can be imaged that, instead of only 3 paths, there are infinitive paths of impedance transformation which can be drawn on the Smith Chart. In other words, there

are infinite choices for the values of the 3 parts in a *Π* type matching network. This therefore provides the designer with a lot of alternatives.

However, such an alternative feature is not available if another restrictive condition is set up. For instance, if the *Q* value of the matching network is one of the design goals, there is only one choice for the value of the parts, C_{P1}, L_S, and C_{P2}.

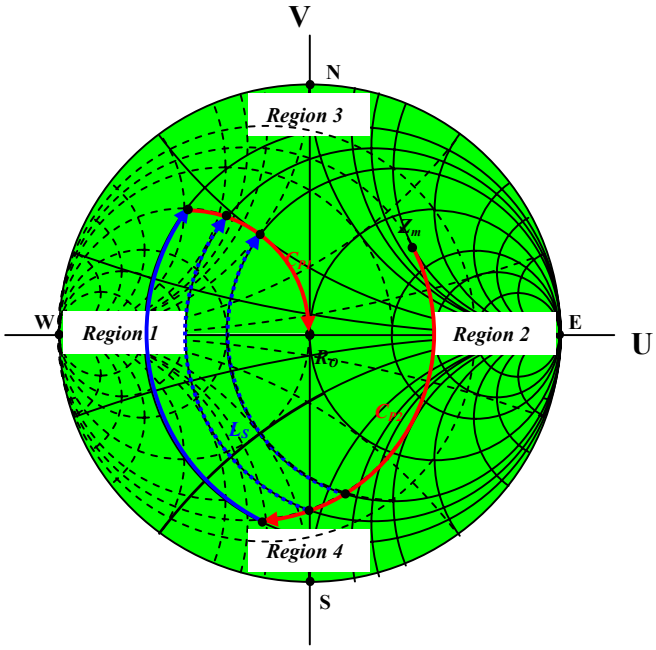


Figure 2.31 Possible moving paths of impedance for the topology C_{p1}-L_S-C_{P2} on the Smith Chart.

2.3.5.3 *T* Type Matching Network

A *T* type matching network contains 3 matching parts. There are 8 possible alternative topologies:

a) C_{S1}-C_P-C_{S2},
b) C_{S1}-C_P-L_{S2},
c) C_{S1}-L_P-C_{S2},
d) C_{S1}-L_P-L_{S2},
e) L_{S1}-C_P-C_{S2},
f) L_{S1}-C_P-L_{S2},
g) L_{S1}-L_P-C_{S2},
h) L_{S1}-L_P-L_{S2}.

Figure 2.32 shows 8 possible topologies of a *T* type matching network and Table 2.2 lists their advantages and disadvantages.

The criteria in the selection of a matching network topology are the same as that for a *Π* type matching network, that is, its DC blocking capabilities and DC short-circuited resistance, and cost. Many features of the topologies are also similar to each other. First, topology a), C_{S1}-C_P-C_{S2}, contains no inductor but 3 capacitors. As we discussed in the

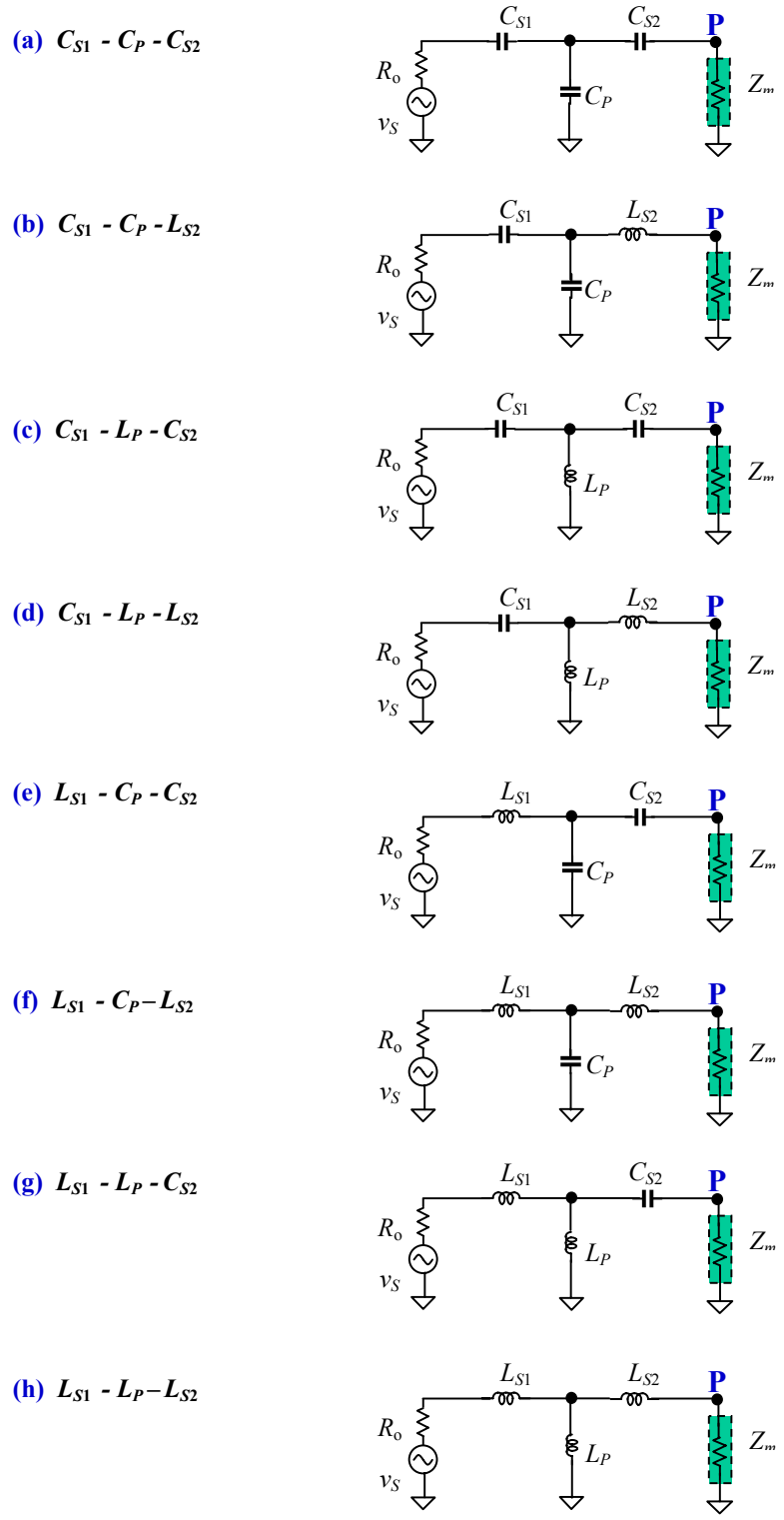


Figure 2.32 Topologies of a *T* type matching network

Table 2.2 Advantages and disadvantages of different topologies in a *T* type matching network

Topology	Inductor number	Advantages	Disadvantages
(a) C_{S1} **-** C_P **-** C_{S2}	0	Lowest cost	Region 3 only
(b) C_{S1} **-** C_P **-** L_{S2}	1	Low cost	
(c) C_{S1} **-** L_P **-** C_{S2}	1	Low cost	
(d) C_{S1} **-** L_P **-** L_{S2}	2		Medium cost DC short-circuited
(e) L_{S1} **-** C_P **-** C_{S2}	1	Low cost	
(f) L_{S1} **-** C_P **-** L_{S2}	2		Medium cost DC un-blocked
(g) L_{S1} **-** L_P **-** C_{S2}	2		Medium cost
(h) L_{S1} **-** L_P **-** L_{S2}	3		Highest cost DC short-circuited DC un-blocked

previous section, it is the most welcome topology since there are many outstanding advantages. However, it should be noted that this topology is only available for the original impedance located in the region 3 of Smith Chart. Secondly, there are 4 topologies, d), f), g), and h), with two or three inductors. Again, they are usually not adapted in the circuit design because they are too expensive.

The remarkable difference between *T* and *Π* type matching networks shows up in those topologies containing only one inductor. In the *T* type matching networks, they are b) C_{S1}-C_P-L_{S2}; c) C_{S1}-L_P-C_{S2}; and e) L_{S1}-C_P-C_{S2}. Contrary to the topologies containing only one inductor in the *Π* type matching network, all of these have no DC short-circuiting or DC blocking problems. From this point, a *T* type network is superior to a *Π* type network.

In a practical design, an appropriate topology must be selected among those topologies that contain one or no inductors. The candidates are a), C_{S1}-C_P-C_{S2}, b), C_{S1}-C_P-L_{S2}, c), C_{S1}-L_P-C_{S2}, and e), L_{S1}-C_P-C_{S2}. The selection, of course, depends on the location of the original impedance Z_m on the Smith Chart because there is a regional prohibition for all the topologies. Figure 2.33 to 2.36 shows their available and prohibited area on the Smith Chat for these 4 topologies.

From Figures 2.33 to 2.36, it can be seen that topology a), C_{S1}-C_P-C_{S2}, can only be applied when the original impedance Z_m is located in region 3. Topology b), C_{S1}-C_P-L_{S2}, can be applied when the original impedance Z_m is located in any region except region 2. Topology c), C_{S1}-L_P-C_{S2}, can be applied for any value of the original impedance Z_m, so it is the easiest to use in the design. Topology e), L_{S1}-C_P-C_{S2}, can be applied when the original impedance Z_m is located in regions 2 and 3.
Usually, the input impedance of a MOSFET device is located in regions 2 or 4, so its input impedance matching network is usually implemented by topology c), while the input impedance of a bipolar is usually located in regions 1 or 3, so its input impedance matching network is usually implemented by topologies b) or c).

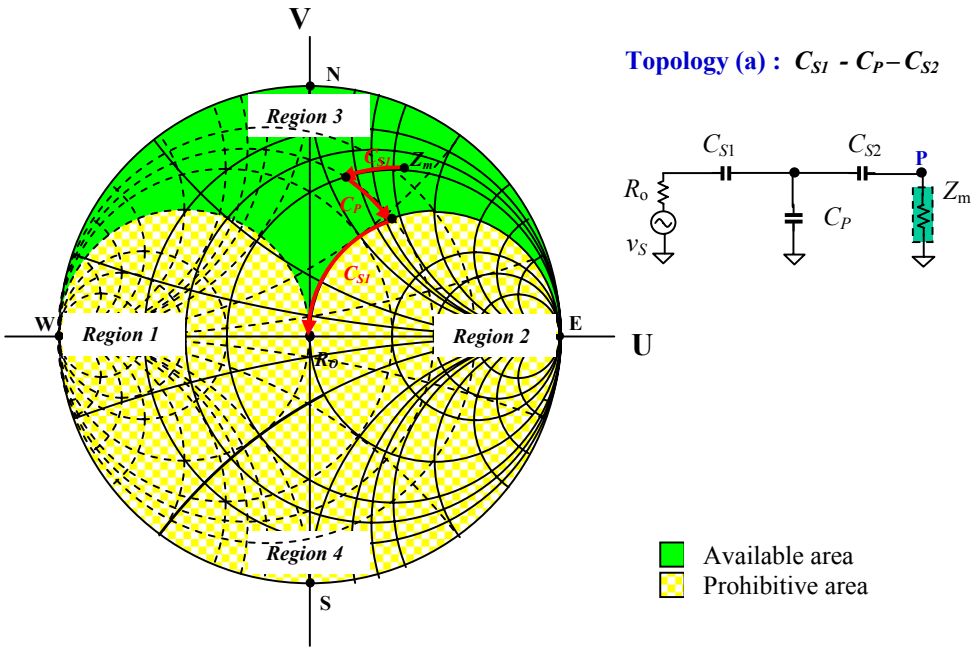


Figure 2.33 Available and prohibitive area of the topology: C_{S1}-C_P-C_{S2}.

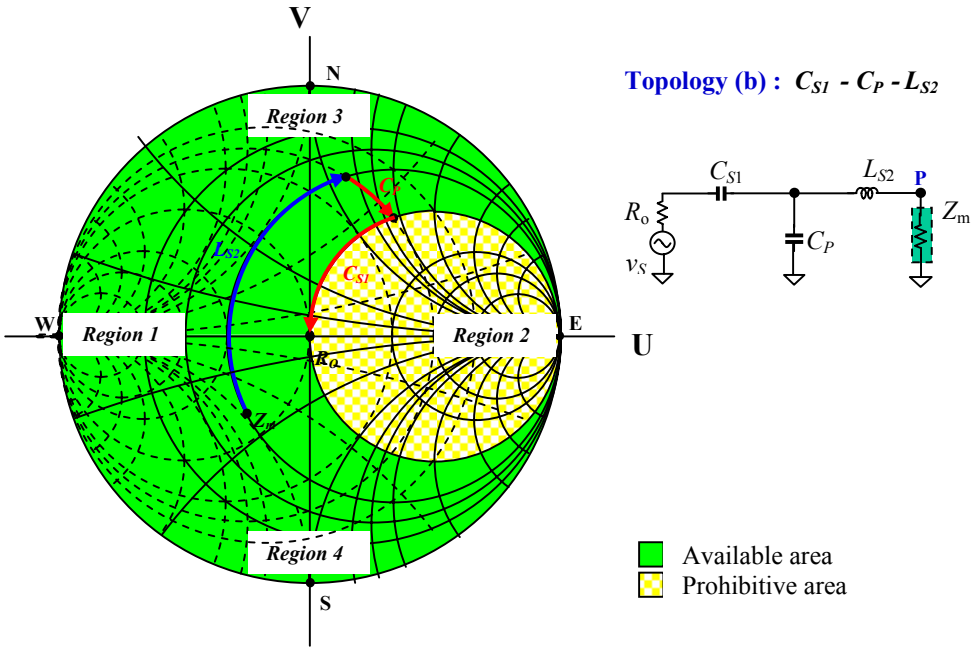


Figure 2.34 Available and prohibitive area of the topology: C_{S1}-C_P-L_{S2}.

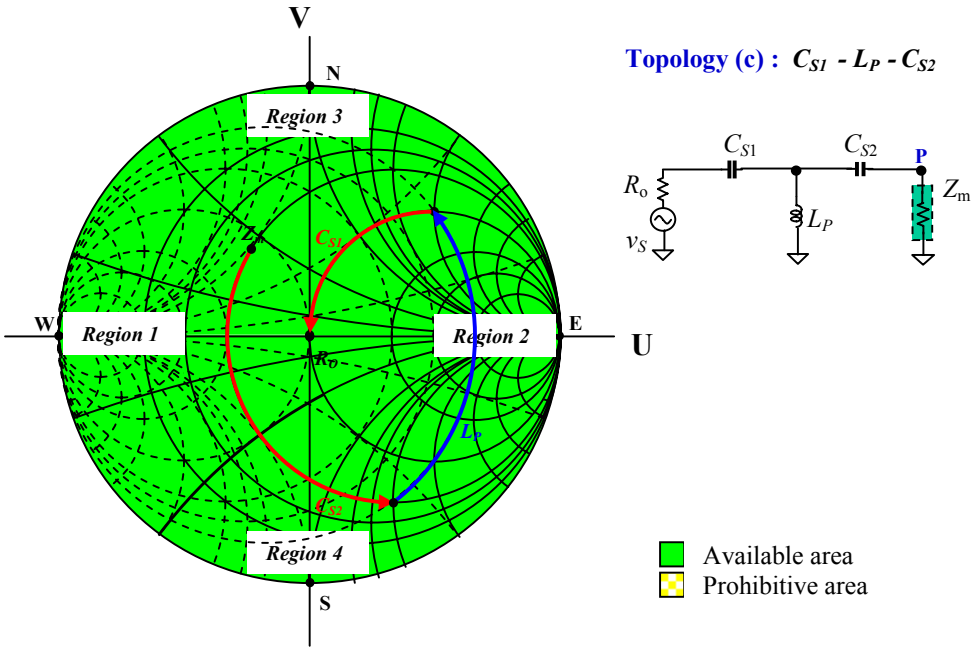


Figure 2.35 Available and prohibitive area of the topology: C_{S1}-L_P-C_{S2}.

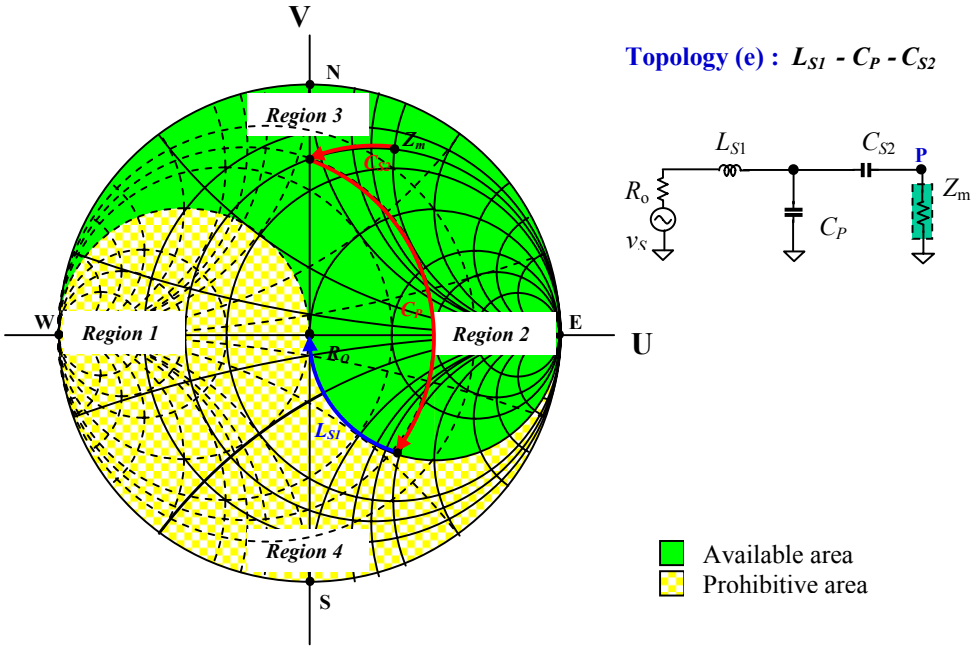


Figure 2.36 Available and prohibitive area of the topology: L_{S1}-C_P-C_{S2}.

Again, let's point out that in a *T* matching network containing 3 parts, there are infinite possible moving paths of the impedance for each topology on the Smith Chart. For example, Figure 2.37 shows three possible moving paths of impedance on the Smith Chart, which corresponds to three possible values of C_{S1}, L_P, *and* C_{S2} for the topology C_{S1}-L_P-C_{S2}. It can be imagined that, instead of only 3 paths, there are infinite paths of impedance transformation which can be drawn on the Smith Chart. In other words, there are infinite choices for the values of the 3 parts in a *T* type matching network. This therefore provides the designer with a lot of alternatives. However, such an alternative feature is no more available if another restrictive condition is set up. For instance, if the *Q* value of the matching network is one of the design goals, there is only one choice for the value of the parts, C_{S1}, L_P, and C_{S2}.

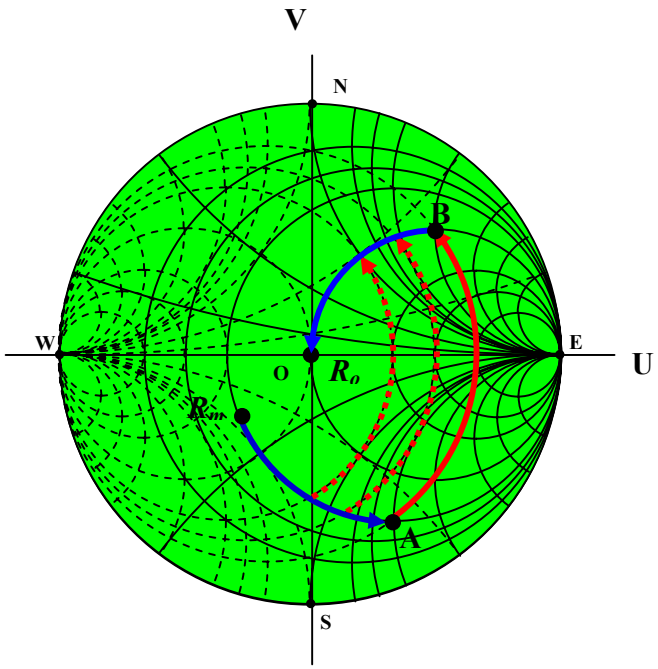


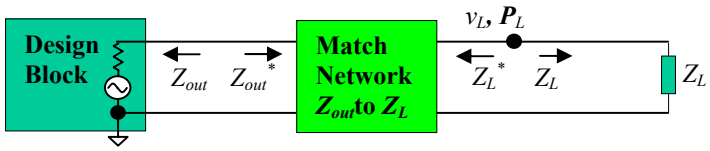
Figure 2.37 *T* type of matching network with topology C_{S1}, L_P, and C_{S2}.

2.4 Some Useful Schemes for Impedance Matching

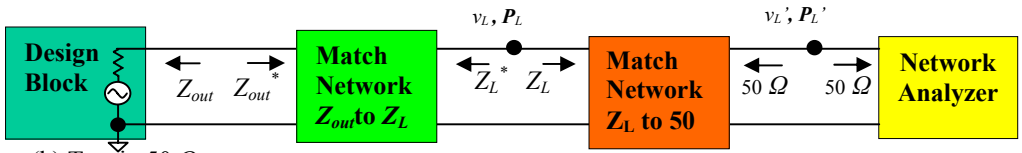
2.4.1 Designs and Tests when Z_L is not 50 Ω

In RF or RFIC designs, it is not necessary to fix the input or the output impedance of the circuit block at 50 Ω because the source of the design block Z_S and the load of the designed block Z_L might be not 50 Ω. Let's discuss only the case when Z_L is not 50 Ω.

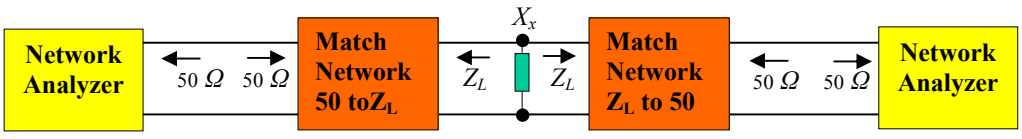
It is no problem to match a designed block to any value of the load Z_L. As shown in Figure 2.38(a), the designed block is matched to Z_L by a match network. The input impedance of the match network is Z_{out}^*, which is to conjugate-match the output impedance of the designed block, Z_{out}, and, its output impedance is Z_L^*, which is to conjugate-match the load of the designed block, Z_L.

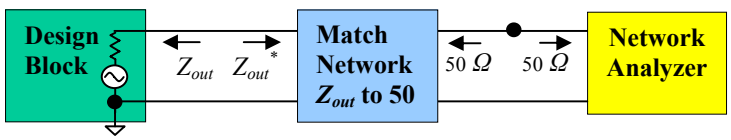


(a) Designed block is matched to Z_L



(b) Test in 50 Ω system





(d) Demonstration in 50 Ω system

Figure 2.38 Design and test when Z_L is not 50 Ω.

The problem is that the testing is only available in a 50 Ω system since the input or output impedance of most test equipments in an RF or RFIC laboratory is 50 Ω. For instance, the power spectrum is often measured by a network analyzer with an input impedance of 50 Ω. In order to make an exact measurement, an additional match network, from Z_L to 50 Ω, must be inserted between the match network from Z_{out}^* to Z_L^* and the network analyzer with its 50 Ω of input impedance as shown in Figure 2.38(b). The additional match network, from Z_L to 50 Ω, will be removed in the actual circuitry application as shown in Figure 2.38(a). Therefore, the actual performance of the design block with its match network, from Z_{out}^* to Z_L^*, must be obtained by the subtraction of the contribution of an additional match network, from Z_L to 50 Ω, from the measurement in the 50 Ω system as shown in Figure 2.38(b). The contribution of an additional match network, from Z_L to 50 Ω, must be measured in the 50 Ω system as well, which is shown in the Figure 2.38(c). It should be noted that there are two identical match networks, from Z_L to 50 Ω, which are piggy-backed together in order to make the exact measurement in a 50 Ω system. Its contribution such as power loss or noise figure, of course, is half of its measured result. In order to reach impedance matching condition for these two piggy-backed blocks, a part with reactance, $X_x=-2X_L$, must be inserted between the piggy-backed sides so as to "neutralize" their reactances existed in two piggy-backed blocks.

An additional design could be worked out as shown in Figure 2.38(d), in which the design block is directly matched to 50 Ω though it is not what we want to apply it into the actual system. The purpose of this design is twofold: One is to demonstrate the possibility to reach the design goals. Another one is to function as a reference for the desired design as shown in Figure 2.38(a).

Summarily, the design and test when Z_L is not 50 Ω must go through four steps as shown in Figure 2.38.

2.4.2 Conversion between "T" and "Π" type Matching Network

If a matching network consists of 3 parts, it can be built with "T" or "Π" configuration as shown in Figure 2.39. It is very hard to say which one is better than the other. Generally, the selection is based on the following considerations:

- Type "T" is better than type "Π" if the DC blocking to the next block is concerned.
- Type "Π" is better than type "T" if the spray reactance in the input of next block is concerned.

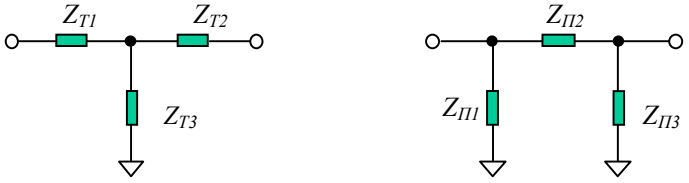


Figure 2.39 "T" and "Π" type of a matching network with 3 parts

As a matter of fact, “T” and “Π” type of block can be converted from each other:

The conversion equations from Π to T are,

$$Z_{T1} = \frac{Z_{\Pi 1} Z_{\Pi 2}}{Z_{\Pi 1} + Z_{\Pi 2} + Z_{\Pi 3}} \quad , \tag{2.85}$$

$$Z_{T2} = \frac{Z_{\Pi 2} Z_{\Pi 3}}{Z_{\Pi 1} + Z_{\Pi 2} + Z_{\Pi 3}} \quad , \tag{2.86}$$

$$Z_{T3} = \frac{Z_{\Pi 3} Z_{\Pi 1}}{Z_{\Pi 1} + Z_{\Pi 2} + Z_{\Pi 3}} \quad . \tag{2.87}$$

The conversion equations from T to Π are,

$$Z_{\Pi 1} = Z_{T3} + Z_{T1} + \frac{Z_{T3} Z_{T1}}{Z_{T2}} \ , \tag{2.88}$$

$$Z_{\Pi 2} = Z_{T1} + Z_{T2} + \frac{Z_{T1} Z_{T2}}{Z_{T3}} \ , \tag{2.89}$$

$$Z_{\Pi 3} = Z_{T2} + Z_{T3} + \frac{Z_{T2} Z_{T3}}{Z_{T1}} \quad . \tag{2.90}$$

“T-Π” conversion is also called “*-Δ” conversion because “T” looks like “*” and “Π” looks like “Δ” as shown in Figure 2.40.

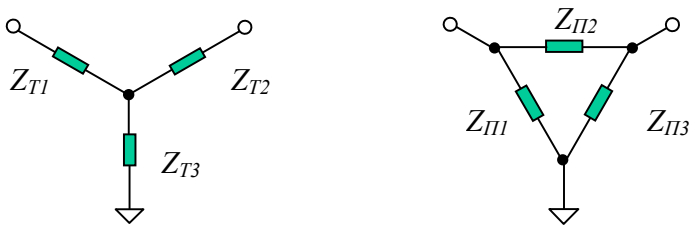
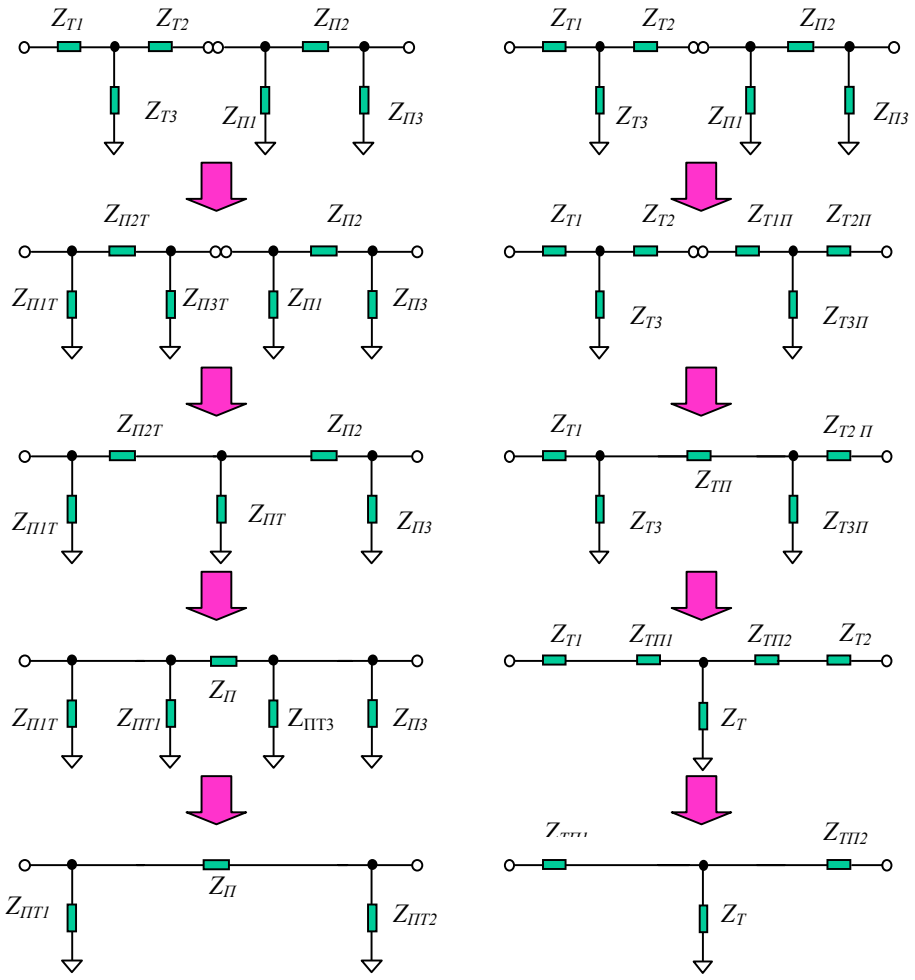


Figure 2.40 “*” and “Δ” type of a matching network with 3 parts.

The example as shown in Figure 2.41 illustrates that the "*T – Π* " or "∗ - *Δ*" conversion provides the flexibility of the topology for a matching network, while its input and output impedance can be kept unchanged. The designer can have 5 choices of topology, in which the part count can be changed from 6 to 3.

The "T – Π " or "∗ - *Δ*" conversion is important to a designer. The reasons are:

- Sometimes the part's value is too low and is comparable with the parasitic parameters. At this point its performance, of course, is nothing to talk about the reliable. On another extreme end, its performance is also not reliable if the part's value is too high.
- Sometimes the designer cannot find the expected value of part in the market. By means of the "T – Π " or "∗ - *Δ*" conversion, the value of part can be changed.
- The total part number can be reduced or increased.



(a) "T" to " Π" (b) " Π" to "T"

Figure 2.41 Conversion and combination of impedance matching networks between "T" and "Π" type

2.4.3 Parts in a Matching Network

In a practical RF or RFIC design, a passive matching network built by one or no part exists scarcely. A passive matching network built by two parts is usually available in most designs though there is a topology limitation. A passive matching network built by three parts provides infinitive choices of the parts' values. Empirically, in order to design a wide-band matching network, more than three parts might be necessary, which scarcely happens in most circuit design. The passive parts in reality are not ideal so that they always attenuate the signal with their resistive component. Too many parts in a matching network might bring about a serious attenuation of signal.

Usually a matching network is built by two kinds of passive parts: capacitor and inductors. We would try to use capacitors as much as possible and try to use inductors as less as possible. The reasons are

- The cost of capacitor is much less than that of inductors if the circuit is built by discrete parts;
- In IC designs, the area of a capacitor is much less than that of an inductor. Therefore, the cost of a capacitor is cheaper than that of an inductor.
- In IC designs, the Q value of an inductor is much lower than that of a capacitor.

2.4.4 Impedance Matching between Power Transportation Units

Nobody would doubt the importance of the impedance matching in an RF or RFIC design. However, there is a paradox about the impedance matching: Is it necessary to do impedance matching from part to part? If yes, how to do it? If no, how come?

To answer this question, we should return to the goals of the impedance matching: 1) To reach the maximum of power transportation; 2) To eliminate the phase shift in the power transportation. If a matching network consists of more than one part, any individual part cannot perform these goals independently. All of parts in the matching network are actually co-operated together to maximize the power transportation and the elimination of phase shift. In other words, a matching network is a basic entity or a "minimum cell" in the performance of the impedance matching goals. Impedance matching between the individual parts is meaningless. It is therefore concluded that the impedance matching is necessary between the power transportation units but not between individual parts except when the matching network consists of only one part.

Very often impedance matching is not conducted between power transportation units in the actual circuit design. For example, in a cascode LNA with CE-CB or CS-CG configuration, the individual units, CE, CB, or CS, CG portion, are the power

transportation units. The impedance matching is usually neglected between CE and CB portion or between the CS and CG portion. This is due to other considerations such as noise or the simplicity of the circuit. Otherwise, it is better to have the impedance matching between CE and CB, or between CS and CG portion if only the power transportation is taken into account.

2.4.5 Impedance Matching for a Mixer

Impedance matching is one of key subjects in an RF circuit design. It is not easy work sometimes. One of the examples is to match a differential mixer as shown in Figure 2.10 or 2.11. Owing to the imperfect isolation between port RF, LO, and IF, the impedance matching work must be conducted with many iterations back and forward between three ports, especially between the LO and the RF port because the isolation between them is poor. The matching procedures should be

1) Matching of the LO port first
   LO injection is the main source to control the operating status of a mixer. The operating status of a mixer never approaches to a normal state until the LO port is well-matched.
2) Matching of the RF port second
3) Repeat the two steps above until the variation between steps is negligible;
   Owing to the imperfect isolation between the port LO and the port RF, the variation resulted from the matching at the RF port would change the matching status at the LO port, and vice versa.
4) Finally, matching of the IF port
   Usually, the isolation between LO and IF, or between RF and IF, is good enough so that the variation resulted from the matching at the IF port will not effect the matching status in both of the LO and RF port.

References

[1] J. Gurley, “Impedance matching by means of nonuniform transmission lines,” Antennas and Propogation, Transactions of the IRE Professional Group on , Vol. 4 , Issue: 1 , December 1952.

[2] T. O'Meara, “Very-Wide-Band Impedance-Matching Network,” Component Parts, IRE Transactions on , Vol. 9 , Issue: 1 , pp. 38 – 44, March 1962.

[3] William H. Hayt, Jr. and Jack E. Kemmerly, “Engineering Circuit Analysis,” (Book), McGraw-Hill, Inc., 1971.

[4] Vendelin, G. D., W. Alexander, and D. Mock, “Computer Analyzes RF Circuits with Generalized Smith Charts,” Electronics, pp.102-109, 1974.

[5] Ralph S. Carson, “High-Frequency Amplifiers,” (Book), John Wiley & Sons, Inc., 1975.

[6] G. J. Laughlin, “A New Impedance-Matched Wide-Band Balun and Magic Tee,” Microwave Theory and Techniques, IEEE Transactions on , Volume: 24 , Issue: 3 , pp.135 – 141, March 1976 .

[7] H. J. Carlin and P. Amstutz, “On Optimum Broad-Band Matching,” IEEE Transactions on Circuits and Systems, Vol. CAS-28, No. 5, pp.401-405, May 1981.

[8] G. P. Young and S. O. Scanlan, “Matching Network Design Studies for Microwave Transistor Amplifiers,” IEEE Transactions on Microwave Theory and Techniques, Vol. MTT-29, No. 10, pp.1027-1035, October 1981.

[9] Guillermo Gonzalez, “Microwave Transistor Amplifiers,” (Book), Prentice-Hall, Inc., 1984.

[10] P.L.D. Abrie, “The Design of Impedance Matching Networks for Radio-Frequency and Microwave Amplifiers,” (Book), Artech House, Norwood, Mass. 1985.

[11] U. L. Rohde, “Designing a Matched Low Noise Amplifier Using CAD Tools,” Microwave Journal, Vol. 29, No. 10, pp.154-160, October, 1986.

[12] K. R. Cioffi, “Broad-band distributed amplifier impedance-matching techniques,” Microwave Theory and Techniques, IEEE Transactions on , Vol. 37 , Issue: 12 , pp.1870 – 1876, December 1989.

[13] George D. Vendelin, Anthony M. Pavio, Ulrich L. Rohde, “Microwave Circuit Design Using Linear and Nonlinear Techniques,” (Book), John Wiley & Sons, Inc., 1990.

[14] J. K. Fidler, Y. Sun, “Computer-aided determination of impedance matching domain,” Digital and Analogue Filters and Filtering Systems, IEE Twelfth Saraga Colloquium on , pp. 1/1 - 1/6, 6 November 1992.

[15] Yichuang Sun, J. K. Fidler, “Design of Π impedance matching networks,” Circuits and Systems, 1994. ISCAS '94., 1994 IEEE International Symposium on , Vol. 5, pp.5 - 8 , 30 May-2 June 1994.

[16] Y. Sun, J. K. Fidler, “Design method for impedance matching networks,” Circuits, Devices and Systems, IEE Proceedings [see also IEE Proceedings G- Circuits, Devices and Systems] , Vol. 143 , Issue: 4 , pp. 186 – 194, August 1996.

[17] Thomas H. Lee, “The Design of CMOS Radio-Frequency Integrated Circuits,” (Book), Cambridge University Press, 1998.

[18] M. Lapinoja, T. Rahkonen, “An active tuning and impedance matching element,” Circuits and Systems, 1998. ISCAS '98. Proceedings of the 1998 IEEE International Symposium on , Vol. 1 , pp. 559 - 562 , 31 May-3 June 1998.

[19] G. Girlando, G. Palmisano, “Noise figure and impedance matching in RF cascode amplifiers,” Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on [see also Circuits and Systems II: Express Briefs, IEEE Transactions on] , Vol. 46 , Issue: 11 , pp.1388 – 1396, November 1999.

[20] B. C. Wadell, “Smith charts are easy. I,” Instrumentation & Measurement Magazine, IEEE , Vol. 2 , Issue: 1 , pp. 37 – 40, March 1999.

[21] B. C. Wadell, “Smith charts are easy. II,” Instrumentation & Measurement Magazine, IEEE , Vol. 2 , Issue: 2 , pp. 45 – 47, June 1999.

[22] B. C. Wadell, “Smith charts are easy. III,” Instrumentation & Measurement Magazine, IEEE , Vol. 2 , Issue: 3 , pp. 38 – 42, September 1999.

Index

Contents

Chapter 3 RF Grounding

3.1 A True Story

There is a true story which took place when I visited an institute in Asia. A senior principal engineer questioned me about the implementation and testing of a power amplifier. For a special purpose he was assigned to build a power amplifier with an output power 200 *W*. He was asked to accomplish the project within one year.

He found the corresponding device from a Motorola catalog and collected all the parts which were available from other companies. Soon he started construction on a PCB (Printed Circuit Board). At the beginning the output power was very weak—in short, far from his goal. He then adjusted the value of all the components, except the device itself, one by one; he ended up applying all the components' values exactly as the Motorola high power amplifier module, which was published in the Motorola product manual. Unfortunately the output power on his tested circuit board was still very poor.

After deliberation and upset for weeks and months, he was ready to believe that the only reason of his failure was due to the difference between his PCB and Motorola's. Accidentally, a friend of his helped him get a footprint of the PCB for high power amplifier from a Motorola published manual. He was very excited and immediately started working on the PCB. "This time, the results may be different!" he thought.

The outcome was so disappointing to him after he rebuilt the power amplifier with the new PCB. The output power was just about 10 *W* while quite a lot of power was found at the other nodes. "What is going on?" He replaced the old devices by new ones again and again but no avail. "Up to now, I have been struggling for a total of 10 months. I tried everything left and right, up and down, but all in vain!"

"Maybe Motorola builds their power amplifiers by good devices and sells the lousy ones in the outside market?" he complained. I told him that I had never found Motorola cheating the customers since the many years I worked for it.

I convinced him to let me check and test the circuit board he is working for. Yes, all the parts he applied on the board were the same as the Motorola's modules and the PCB were almost identical to each other as well. Yes, quite a lot of power appearing at other nodes but not at the output… Eventually I found the strongest power appearing along the DC power supply runner!

"You didn't do the grounding well, my dear friend!" I said. But, he disagreed with me and showed me what he had done for the grounding: "Look! I applied 8 capacitors with different values, from 10 *pF* to 10 *μF*, connected from DC power supply to the ground. How come it did not function well with the short-circuited AC?"

I did see that he connected all of 8 capacitors between the runner of the DC power supply and the ground pad as shown in Figure 3.1. The first group of capacitors, from

10 *pF* to 500 *pF,* were chip capacitors while the second group of capacitors, from 1000 *pF* to 10 *μF*, were electrolytic capacitors.

I told him that it was a problem due to faulty RF grounding. No capacitor is appropriate for an AC short-circuited.

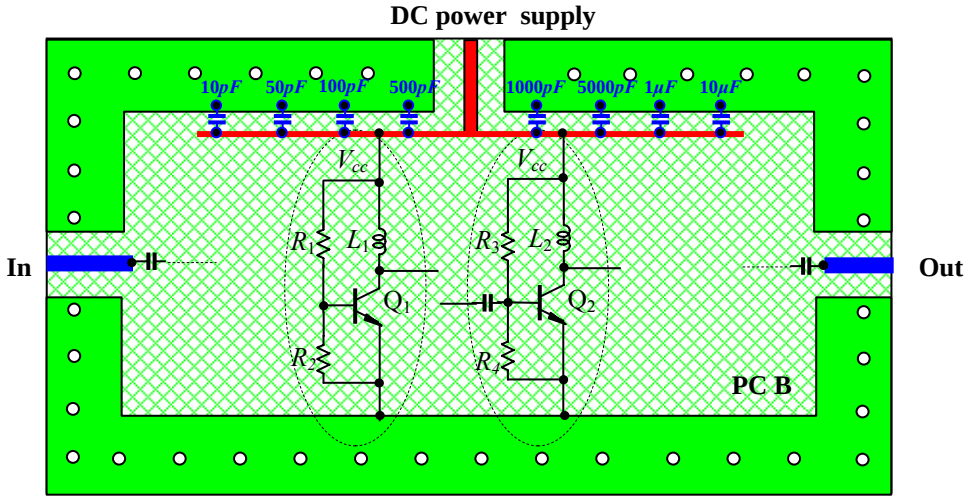


Figure 3.1 "Possible" RF grounding by capacitors

Top metallic area
Bottom metallic area
Conductive via from top to bottom
Capacitor

We then took away all 8 capacitors between the DC power supply runner and the ground. Instead, a chip capacitor was chosen from their engineering stock room and connected between the DC power supply runner and the ground nearby the input of DC power supply. After doing this, the RF power disappeared from the DC power supply runner and, thanks God, the power amplifier started to work in a normal operating state.

"There has been continuous trouble for 10 months. Now it is gone after 10 minutes' work," he concluded. In that night, the senior principle engineer slept peacefully, I believe.

The purpose of this true story is to illustrate how important RF grounding in the RF circuit design is. We are going to discuss it in a somehow more detailed manner in this chapter.

3.2 Three Components for RF Grounding

On an RF circuit board, a common RF grounding point must be provided as a "zero" or a reference point to all the RF signals at any input and output port. The "common" means that these points must be equipotential. At another specific points on an RF circuit board, such as the terminals of DC power supply and DC bias, the DC voltage must be kept at the original value of the DC power supply but the RF signal must be short-circuited to the ground or rejected to a level much lower than conceivable from any RF source. In other words, at the terminals of the DC power supply and the DC bias, the impedance must be infinitive to the DC current or voltage but must approach to zero to the AC or RF current or voltage. These are the objectives of RF grounding.

RF grounding on a circuit board is an indispensable segment in every RF block, no matter in passive or active circuitry. Imperfect RF grounding results in incorrect measurements for various parameters in the testing of a circuit. Furthermore, the imperfect RF grounding degrades the performance of a circuitry, such as additional noise and spurious, undesired coupling or interference, poor isolation between blocks or parts, additional power loss or radiation, additional phase shift, and at the extreme end, performance with unexpected erroneous functions.

There are three kinds of components applied in RF grounding: "zero" capacitors, micro strip lines, and RF cables, while "infinitive" inductors are good assistants. By means of these components, the task of RF grounding can be accomplished appropriately.

3.2.1 "Zero" Capacitors

Let's assume that a certain point *P* on a runner, as shown in Figure 3.2, is to be grounded. Also, let's assume that the PCB (Printed Circuit Board) shown in Figure 3.2 has two metal layers: the top and bottom metallic layer. The entire bottom metallic layer is the grounding area. The metallic rectangular frame on the top is another portion of grounding area which is connected with the entire bottom metallic layer by many conductive vias (holes). The detailed circuitry with parts is not drawn in Figure 3.2 because we are only focusing on the discussion of RF grounding.

Ideally speaking, point *P* would be grounded perfectly if a special component with "zero" impedance is connected between *P* and *G*, a well grounded point. It is well known that a by-pass capacitor with "zero" impedance is such a special component and is called the "zero" capacitor. The key point now is how to make the impedance of a capacitor actually approach to "zero."

Theoretically the impedance for an ideal capacitor, Z_c, is

$$Z_c = \frac{1}{jC\omega} \quad , \tag{3.1}$$

where C = Capacitance of the capacitor,
ω = Operating angular frequency.

It can be seen that the impedance of a capacitor, Z_c, is infinitive to the DC current or voltage when $\omega = 0$ and that it diminishes as the value of the capacitor increases when $\omega \neq 0$. Ideally its RF impedance could approach to zero by infinitively increasing the capacitor's value C. Thus, RF grounding could be done by connecting an ideal capacitor with infinitive capacitance between the desired RF grounded terminal and a real grounded point. Generally, infinitive capacitance is not realistic or practical. As long as its capacitance is high enough that the RF signals can be short-circuited to the ground below a conceivable level, it is fine.

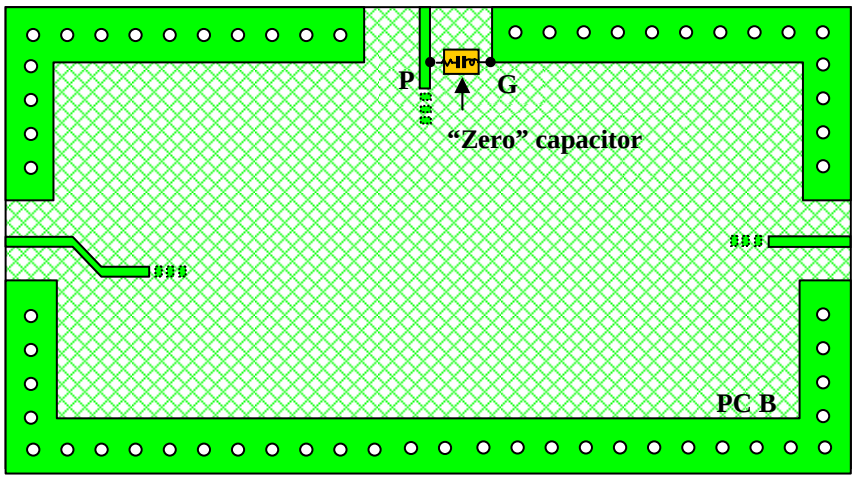


Figure 3.2 RF or AC grounded by a "zero" capacitor on a PCB (Printed Circuit Board).

Top metallic area; P : Point to be grounded;
Bottom metallic area; G : Well-grounded point.
Conductive via from top to bottom;
"Zero" capacitor

The problem is that an ideal capacitor is never available though the capacitors in the low frequency range are pretty close to being ideal and can be described by expression (3.1). In the RF range, a practical capacitor is never close to being an ideal capacitor so that the ideal formula for the capacitor impedance is even less available. A practical capacitor in an RF range has additional inductance and resistance in series. Instead of infinitively increasing of the capacitor's value C, the impedance of a practical capacitor, Z_c, can approach to zero through its self-resonance in series when its additional inductance is resonanting with the capacitance itself at its SRF (Self-Resonant Frequency).

A by-pass capacitor has an inherent SRF. Its function is to make the DC impedance infinitive but the AC or RF impedance zero at operating frequency. A special by-pass capacitor is chosen so that its SRF is identical with the operating frequency in the circuit. Eventually the "zero" capacitor is re-recognized as a by-pass capacitor with "zero impedance" at its SRF or operating frequency.

There are many kinds of capacitors available in the market. With the advantages of a small size, cheaper price, and simplicity, chip capacitors have become popular electronic parts at present. There will be a more detailed discussion on the "zero" chip capacitor later, in chapter 4.

It should be noted that the "infinitive" inductor is a good assistant to the "zero" capacitor which plays the main role in the RF grounding. Theoretically the impedance of an ideal inductor**,** Z_L, is

$$Z_L = jL\omega \quad , \tag{3.2}$$

where L = Inductance of the inductor,
ω = Operating angular frequency.

It can be seen that the impedance of an inductor, Z_L, is zero to the DC current or voltage when $\omega = 0$ and that it is increased as the value of the inductor is increased when $\omega \neq 0$. The value of Z_L can be increased to a very high level, almost infinitive, if the value of the inductor, L, is high enough for the RF signal.

However, just like the capacitor, an ideal inductor is never available, though a real inductor in the low frequency range is pretty close to be ideal and can be described by expression (3.2). In an RF range, this ideal formula for the inductor impedance is even less available. Instead, a practical chip inductor in the RF range has additional capacitance and resistance in parallel. The impedance of a practical chip inductor becomes very high when its additional capacitor is resonanting with the inductance itself in parallel at an operating frequency. It is called an "infinitive" inductor because ideally, its impedance approaches infinity at its SRF.

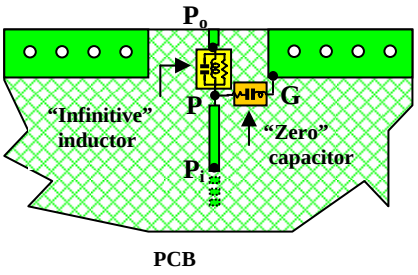
Figure 3.3 shows how an "infinitive" inductor can be very helpful to the RF grounding. To save space, Figure 3.3 duplicates only the portion containing the point to be RF grounded, P. It shows that there are three ways to apply the "infinitive inductors" to assist RF grounding. In Figure 3.3(a), an "infinitive" inductor is inserted between point P_o and point P, where P_o is the outside adjacent point on the same runner before the "infinitive" inductor is inserted. This "infinitive" inductor blocks the RF signal from outside point P_o to the point P. At point P, the voltage or power of RF signal from point P_o would be significantly reduced to below a conceivable level but the DC voltage at point P is kept the same as at point P_o, since the impedance of the "infinitive" inductor is approaching to infinitive. Point P is in fact very close to being RF grounded even if the "zero" capacitor is imperfect, as long as the point P_o is the unique RF source.

Similar to the case shown in Figure 3.3(a), in Figure 3.3(b), an "infinitive" inductor is inserted between point P and P_i, where P_i is the inside adjacent point on the same runner before the "infinitive" inductor is inserted. This "infinitive" inductor blocks RF signals from inside point P_i, to the point P. At point P, the voltage or power of RF signal from point P_i would be significantly reduced to below a conceivable level but the DC voltage at point P is kept the same as at point P_i, since the impedance of the "infinitive" inductor is "infinitive". Point P is actually very close to be nicely RF grounded even if the "zero" capacitor is imperfect, as long as the point P_i is the unique RF source.

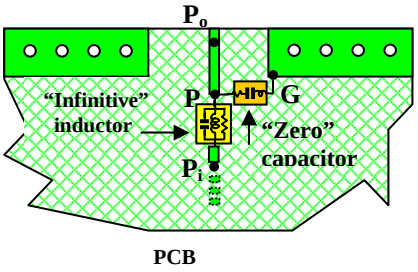
Finally, Figure 3.3(c) shows the combination of the above technologies as shown in Figure 3.3(a) and (b). It, of course, is the best from the viewpoint of RF grounding though it is rather costly.

In the consideration of RF grounding, it should be noted that the "infinitive" inductor must be chosen so that its SRF is identical with the operating frequency in the circuit.

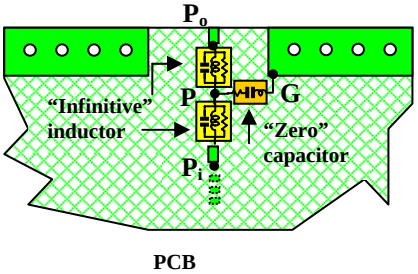
Historically this "infinitive" inductor was also called the "RF choke".



(a) RF grounding at point P by a "zero" capacitor and the assistance of one "infinitive" inductor ("Infinitive" inductor is connected between P_o and P.)

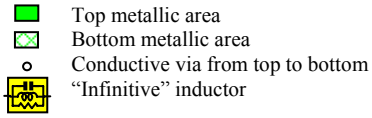


(b) RF grounding at point P by a "zero" capacitor and by the assistance of one "infinitive" inductor ("Infinitive" inductor is connected between P and P_i.)



(c) RF grounding at point P by a "zero" capacitor and by the assistance of two "infinitive" inductors.

Figure 3.3 Three ways of RF grounding by "zero" capacitor and by the assistance of chip inductor on a PCB.



With the advantages of a small size and simplicity, chip inductors have become popular electronic parts at present though it is much more expensive than chip capacitors.

There will be a more detailed discussion on the "infinitive" chip inductor in chapter 4.

3.2.2 Micro Strip Line

If a point to be grounded on a circuit board is not only for the AC signals but also for a DC voltage or current, the usual way is to connect this point to a well grounded point by a metallic runner with high conductivity, such as a copper runner with gold-plating on a PCB or a gold runner on the integrated circuit. This is a unanimous means of grounding when the circuit is operating in the low frequency range. However, it might be unfeasible when the circuit is operating in the RF range. Essentially, a metallic runner with high conductivity either on the IC substrate or on the PCB in the RF range is a micro strip line as shown in Figure 3.4. There is a distributed capacitance and inductance to the ground plane for a PCB or to the substrate for the integrated circuit. Also, there is a distributed resistance due to the imperfect metal conductivity and distributed conductance due to the leakage. Consequently, unless special treatment is made, a metallic runner might not be a good connector for most cases of RF grounding though its conductivity is high enough so that its resistance is negligible.

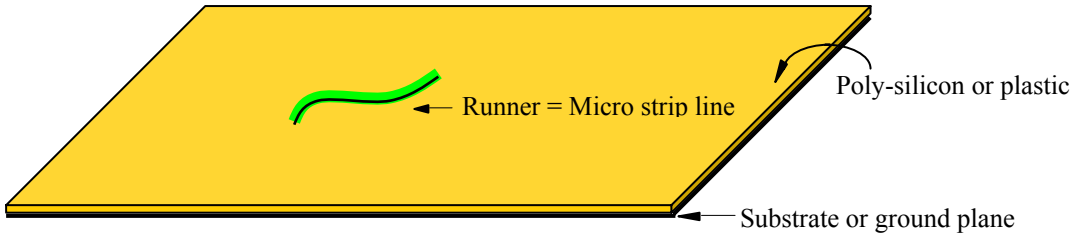


Figure 3.4 A runner on substrate or grounded plane is a micro strip line

Based on the transmission line theory, a micro strip line can be characterized by its distributed parameters. Figure 3.5 shows a runner with a length of l connected between point A and B. Its distributed parameters are: the resistance per unit length, R_S, the inductance per unit length, L_S, the conductance per unit length, G_P, and the capacitance per unit length, C_P. Thus, its characteristic impedance, Z_o, is

$$Z_o = \sqrt{\frac{R_S + jL_S\omega}{G_P + jC_P\omega}} \quad . \tag{3.3}$$

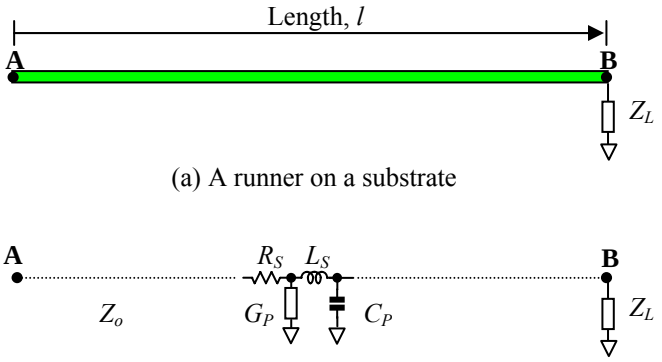
Looking from point A towards point B, the impedance at point A, Z_A, is no longer equal to Z_L but is a function of Z_o, Z_L, γ and l, that is,

$$Z_A = Z_o \frac{Z_L \cosh \gamma l + Z_o \sinh \gamma l}{Z_o \cosh \gamma l + Z_L \sinh \gamma l} \quad , \tag{3.4}$$

where γ is the transmission coefficient and can be expressed by

$$\gamma = \sqrt{(R_S + jL_S\omega)(G_P + jC_P\omega)} \quad , \tag{3.5}$$

and ω is the operating angular frequency.



(a) A runner on a substrate

(b) The equivalent micro strip line of a runner

Figure 3.5 A runner on a substrate and its equivalent micro strip line when B is loaded with Z_L.

For a low loss micro strip line,

$$R_S << L_S\omega \quad , \tag{3.6}$$

$$G_P << C_P\omega \quad , \tag{3.7}$$

and thus expressions (3.5) and (3.4) can be simplified as

$$\gamma = j\omega\sqrt{L_S C_P} = j\beta \quad , \tag{3.8}$$

and,

$$Z_A = Z_o \frac{Z_L \cos\beta l + jZ_o \sin\beta l}{Z_o \cos\beta l + jZ_L \sin\beta l} \quad , \tag{3.9}$$

where

$$\beta = \omega\sqrt{L_S C_P} \quad . \tag{3.10}$$

The wavelength λ is

$$\lambda = \frac{2\pi}{\beta} = \frac{2\pi}{\omega\sqrt{L_S C_P}} \quad , \tag{3.11}$$

when $\beta l = 2\pi$.

All of these distributed parameters are complicated functions of the size, shape and thickness of metal of the micro strip line and its environment.

Now let's find out the conditions when point A is actually grounded.

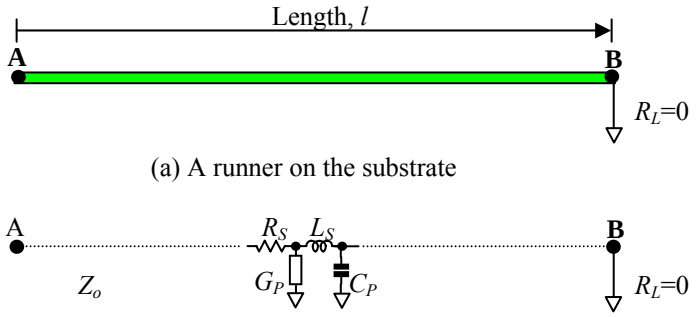
1) In the case of point B being a real RF grounding point, that is,

$$Z_L = 0 \quad . \tag{3.12}$$

In this special case, Figure 3.5 becomes Figure 3.6. If we would like point A to be grounded, the first thing to do is to simply connect point A to point B by a runner with a length of *l*. Point A is actually grounded only when the length of *l* is taken with some special values.

On the basis of expression (3.9), the impedance at point A is

$$Z_A = Z_o j \tan \beta \cdot l \quad . \tag{3.13}$$



(a) A runner on the substrate

(b) the equivalent micro strip line

Figure 3.6 A runner on substrate and its equivalent micro strip line when B is a real RF grounded point, that is, R_L=0

Equation (3.13) indicates that

If
$$0 \le \beta l \le \frac{\pi}{2} \ , \tag{3.14}$$

or

$$0 \le l \le \frac{\lambda}{4} \ , \tag{3.15}$$

then

$$0 \le Z_A \le \infty \quad . \tag{3.16}$$

Z_A is inductive.

The conditions that point A can be grounded from point B therefore are,

a) If

$$l << \frac{\lambda}{4} \ , \tag{3.17}$$

then,

$$Z_A \to 0 \quad . \tag{3.18}$$

b) If

$$\beta l = n\pi \ , \tag{3.19}$$

where n = 0,1,2,3,4…

or

$$l = \frac{n\pi}{\beta} = \frac{n\pi}{\omega\sqrt{L_S C_P}} = n\frac{\lambda}{2} \ , \tag{3.20}$$

then,

$$Z_A = 0 \ . \tag{3.21}$$

It is shown in Figure 3.7.

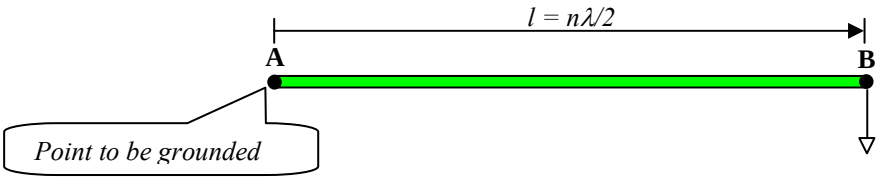


Figure 3.7 Grounding of point A by means of a $n\lambda/2$ micro strip line.

It is then concluded that in order to ground point A, a simple connection by a conductive metallic runner from point A to the real grounded point B will fail unless its length is either greatly less than the quarter wavelength as shown in equation (3.17) or a multiple of its half wavelength as shown in equation (3.20). With the exception of the case as expressed by (3.17), the shortest length of the micro strip line to ground point A when Z_L=0 is

$$n = 1, \tag{3.22}$$

then, from (3.20),

$$l = \frac{\lambda}{2} \ , \tag{3.23}$$

It should be noted that, strictly speaking, the grounding condition (3.20) is correct only for one frequency. However, it is a good approximation within a bandwidth as long as the bandwidth is not too wide.

2) In the case of point B being a real RF open-circuit, that is,

$$Z_L \to \infty \ , \tag{3.24}$$

In this special case, Figure 3.5 becomes Figure 3.8.

If the length of l is taken with some special values,

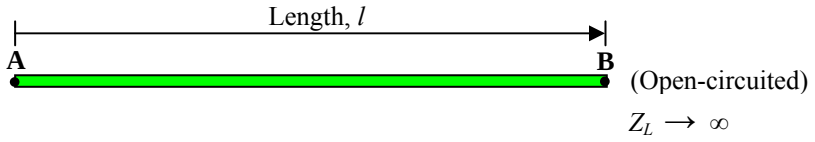
$$l = (2n+1)\frac{\lambda}{4}, \tag{3.25}$$

where n=0,1,2,3,4…

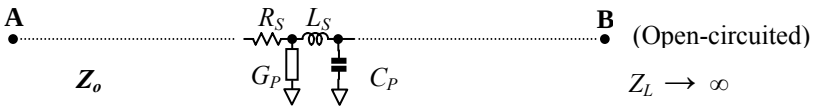
Under the conditions of (3.24) and (3.25), expression (3.9) becomes

$$Z_A = \frac{Z_o^{\ 2}}{Z_L} \to 0 \quad , \tag{3.26}$$

which indicates that the point A is effectively grounded.



(a) A runner on the substrate



(b) Equivalent micro strip line

Figure 3.8 A runner on substrate and its equivalent micro strip line when B is a real RF open-circuited, that is, $Z_L \to \infty$.

It is therefore concluded that in the case of an open-circuited point B, point A, is a grounded point when the length of a micro strip line connected between B and A is the odd multiple of its quarter wavelength as shown in equation (3.25). The shortest length of a micro strip line to point A when $Z_L \to \infty$ is

$$n = 0\,, \tag{3.27}$$

then, from (3.25), we have

$$l = \frac{\lambda}{4} \quad , \tag{3.28}$$

which is shown in Figure 3.9.

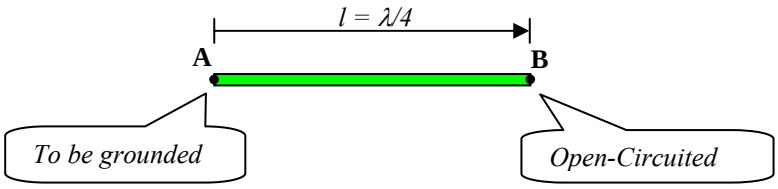


Figure 3.9 Grounding of point A by means of a $\lambda/4$ micro strip line when point B is opened.

A quarter wavelength is shorter than a half wavelength. This is the obvious advantage of RF grounding by a quarter wavelength than by a half wavelength of a micro strip line. However, it should be noted that at open-circuited point B the infinitive

impedance is just a theoretical approximation. Instead, some spray capacitance always exists, especially in a high RF range.

3.2.3 RF Cables

An RF cable is not suitable to be applied in the current RF circuit blocks due to its unacceptably big size. However, it can sometimes be employed to connect equipment in a test set-up.

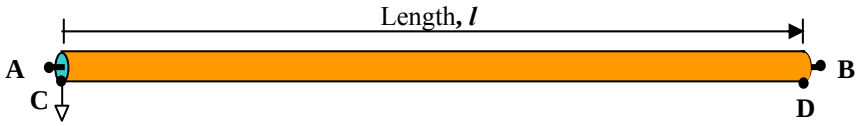


Figure 3.10 Two ends of an RF coaxial cable

Figure 3.10 is a representation of a coaxial cable. A cylindrical conductor with a smaller diameter is its central axis. A coaxial cylindrical conductor with a bigger diameter forms a shielded cavity. The isolated material with high electric permittivity is filled between the central axis and the outer cylindrical conductor. This kind of RF coaxial cable is widely applied in the connection between test equipment, the unit to be tested, the RF input and output terminals. The power of RF signal can be transported from one end of the cable, say, point A, to the other end, say, point B, with negligible attenuation. Let us now check out an interesting question: Assuming that point C is a point at end A on the outer cylindrical conductor and is a real grounded point, is point D, at end B on the outer cylindrical conductor, also a grounded point? In order to answer this question, let's introduce the configuration of a cable balun.

A cable balun is formed by a simple cable with a quarter wavelength as shown in Figure 3.11. End A is a single-ended terminal and point C is a real RF-grounded point. At the other end, point B and D are a good pair of differential terminals, given the length of the cable is a quarter wavelength. Point D is not a grounded terminal though it is on the outer cylindrical conductor.

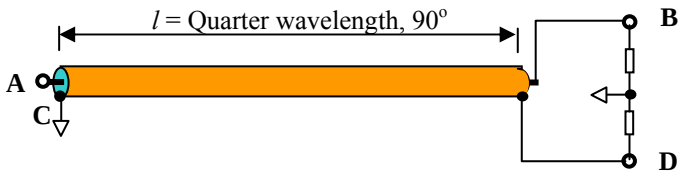


Figure 3.11 The cable balun is a cable with quarter wavelength.

A very important conception follows. In the DC or low frequency range, a metallic surface or plane made by material with high conductivity is an identical potential surface or plane, on which the AC voltage is almost the same everywhere. This surface or plane is an equipotentially-grounded surface or plane when any point on

this surface or plane is connected to the ground. In an RF range, however, a metallic surface or plane made by material with high conductivity is not an identical potential surface or plane, on which the AC voltage is not the same in general. This surface or plane is not a grounded surface or plane even if a point, with the exception of some special points, on this surface or plane is connected to the ground.

Let's return to the subject of the RF cable. Figure 3.12 shows an RF cable connected between an RF generator and an RF power meter. The power of the RF signal from an RF generator is connected to end A of the cable. Point G_g is the grounded point of the RF generator and is connected to point C, which is the end-point on the outer cylindrical conductor at end A. At the B end of the cable, point B is connected to the input of the RF power meter. Point G_p is the grounded point of the RF power meter and is connected to point D, which is the end-point on the outer cylindrical conductor at end B.

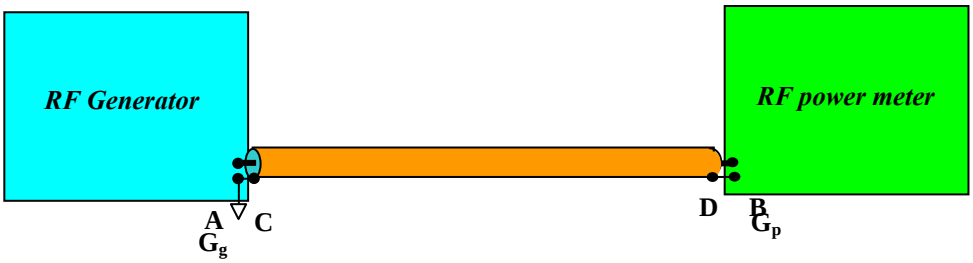


Figure 3.12 A RF cable connected between a RF generator and a RF power meter.
G_g – Ground of the RF generator;
G_p – Ground of the RF power meter.

Up to the argument above, the ground of the RF generator, G_g, is not the same as the ground of the RF power meter, G_p, though the power of the RF signal is transported to the end B and can be measured precisely between B and D. Point D, or G_p, can have the same grounding equipotential as point C, or G_g, only in some special cases when

$$l = n\frac{\lambda}{2} \quad , \tag{3.29}$$

where n = 1,2,3,4…

3.3 Examples of RF Grounding

3.3.1 Test PCB

In order to test an RF or RFIC block or system correctly, the first thing is to design the test PCB well.

There are 3 ways to test an RFIC:
1) Directly test the RFIC on the wafer in an IC probe station.
   The IC die could be either an individual block, multiple blocks, or a SOC (System On Chip) system;
2) Cut the block or system from a wafer as an IC die first, and then bond the IC die on a test PCB for testing.
   This is called COB(Chip On Board) testing;
3) Cut the block or system from a wafer as an IC die first, package it secondly, and finally put it as a part on the test PCB for testing.
   This is called POB(Package On Board) testing;

A test PCB is not necessary for the first approach, directly testing on the wafer, but is necessary for the second and third techniques, COB and POB testing, as shown in Figures 3.13 and 3.14.

And, there are 2 ways to test an RF module:
1) Treat the RF module as a part, put it on the test PCB and test it;
2) Directly build an RF module on the test PCB by using discrete parts.

Figure 3.15 and 3.16 shows the test PCB for testing of RF module as a part and RF module built by discrete parts respectively.

Before the design work of a test PCB is started, the following two items must be seriously considered:

1) Operating frequency of PCB.
   The upper operating frequency of a test PCB is limited by the attenuation of the signal which is operated or transported on the PCB. The attenuation factor is mainly determined by the material of the board. For example, the material of a *FR4*-type of PCB does not have an unacceptable amount of attenuation for a given operating frequency up to 3 to 4 *GHz*. However, it becomes a very lossy medium if the operating frequency gets any higher than about 5 *GHz*. Roughly the working PCB must guarantee its highest operating frequency at least 3 times that of the operating frequency. Otherwise, unacceptable attenuation will occur.

2) Size of PCB.
   The means of grounding a PCB depends on the size of the PCB. If the size of the test PCB is much less than the quarter wavelength of a runner, that is,

$$L << \lambda/4 , \tag{3.30}$$

   where L = the dimension of the PCB, and

λ = Electric wavelength of a runner on the PCB, when the characteristic impedance of the runner, Z_o, is 50 Ω.

Expression (3.30) is the definition of a small-sized PCB electrically. On the contrary, if the size of the test PCB is about or greater than the quarter wavelength of a runner, that is,

$$L \sim \text{or} > \lambda/4 \ . \qquad (3.31)$$

Expression (3.31) is the definition of a large-sized PCB electrically. We will discuss these two cases in the following two sub-sections.

It should be noted that the electric wavelength of a runner on the PCB, λ, depends on the width of the runner or its related characteristic impedance, Z_o, since a runner on the PCB is a segment of a micro strip line. It is quite reasonable to take 50 Ω as a standard characteristic impedance of the runner because in the actual testing the input and output is usually connected to the equipments with 50 Ω of the impedance. By taking 50 Ω as a standard reference, the electric wavelength of a runner on the PCB, λ, can be calculated on the basis of the electric medium constant or the permittivity, the permeability, and the thickness of the PCB.

In the testing of an individual block, such as LNA, Mixer, VCO and so on, the condition of (3.30) is usually satisfied in both of RF module or RFIC circuit design. In the testing of a system, both conditions (3.30) and (3.31) are possible. However, we would always try to use small size of PCB as possible because condition (3.30) ensures the metallic grounded area to be a good equipotential surface.

3.3.1.1 Small Test PCB

3.3.1.1.1 Basic Types of Test PCB

A test PCB with a small size might contain one or more of the 4 basic types of RFIC or RF modules listed below:

1) COB (Chip On Board);
2) POB (Package On Board);
3) RF module as one part on board;
4) RF module directly built on board by discrete parts.

For convenience, a test PCB with small size is supposed to contain only one of the 4 basic types of RFIC or RF module in the following discussion.

Figures 3.13 to 3.16 illustrate the main points in the design of a test PCB containing only one type of RF ICs or RF modules. The test Board that we discuss here is a PCB with 2 metal layers on top and bottom side, which is a very popular applied design in the engineering world today. The test PCBs as shown in Figure 3.13 and 3.14 contain RFIC dies while the test PCB as shown in Figure 3.15 and 3.16 contain RF modules.

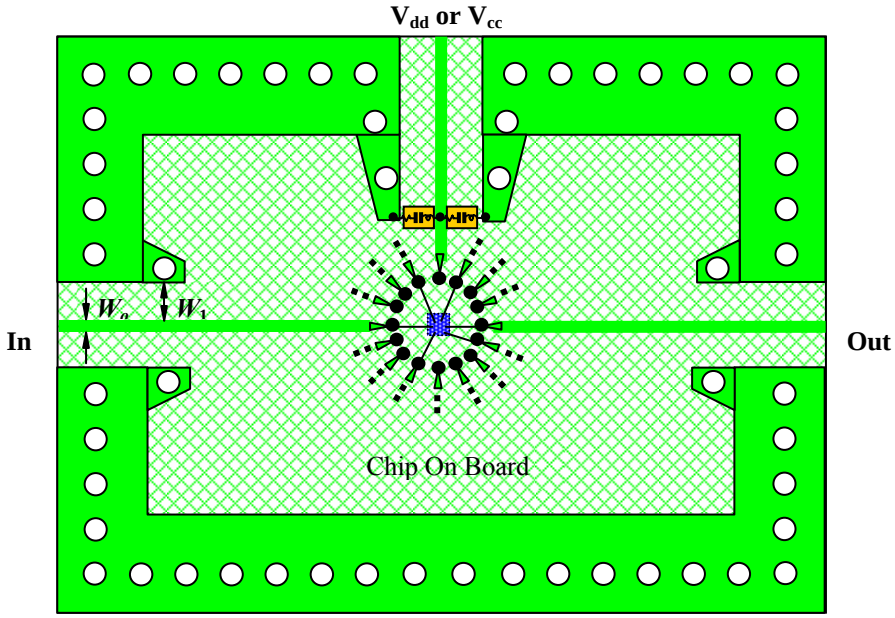


Figure 3.13 Layout of a tested PCB for RFIC COB testing
with 50 Ω characteristic impedance of input and output runners.

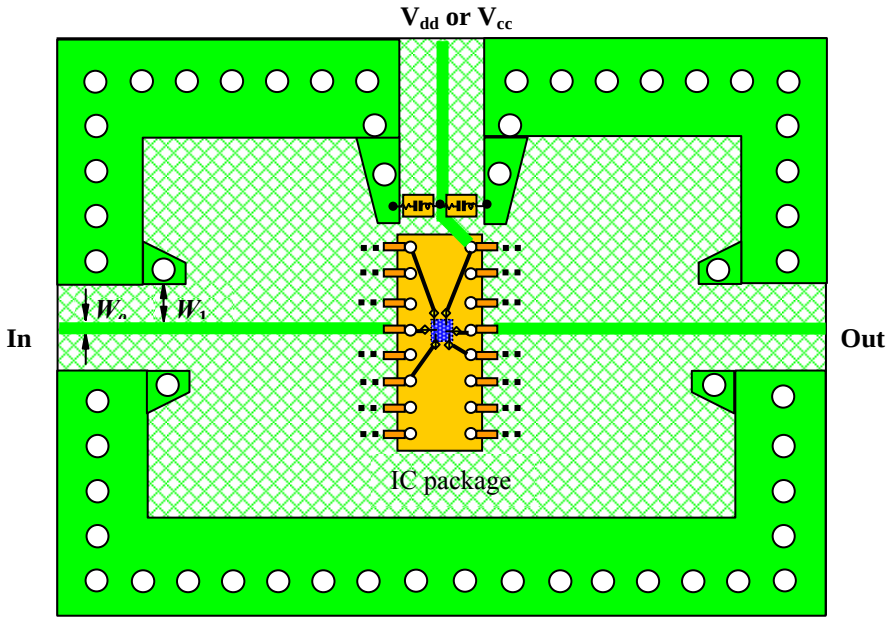
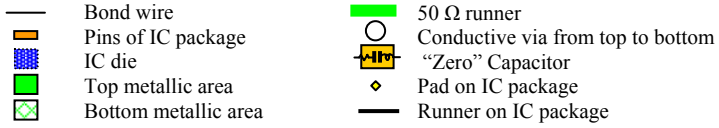


Figure 3.14 Layout of a tested PCB for RFIC POB testing
with 50 Ω characteristic impedance of input and output runners.



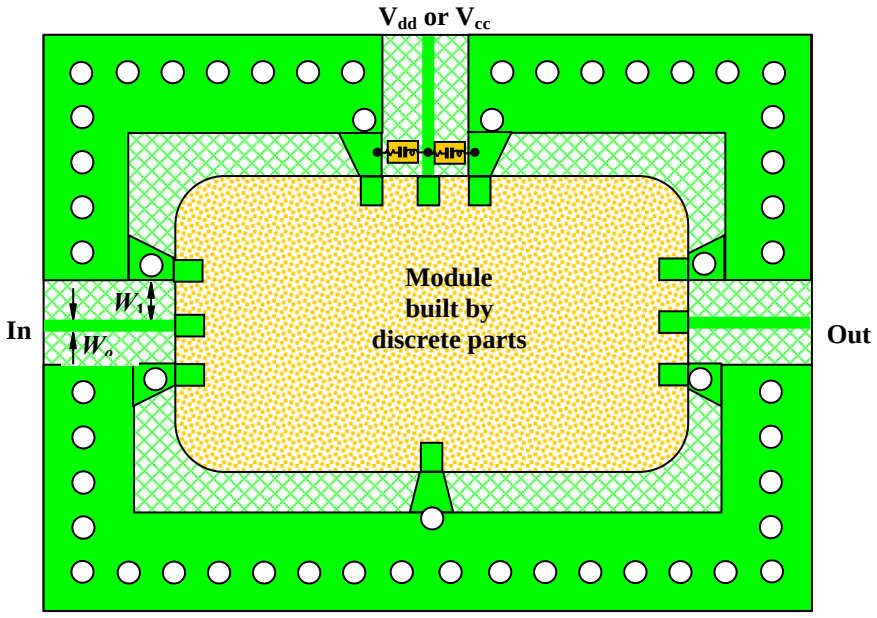


Figure 3.15 Layout of a tested PCB for an RF module as a part, with 50 Ω characteristic impedance of input and output runners.

Top metallic area
Bottom metallic area
Module built by discrete parts
50 Ω runner
Conductive via from top to bottom
"Zero" Capacitor

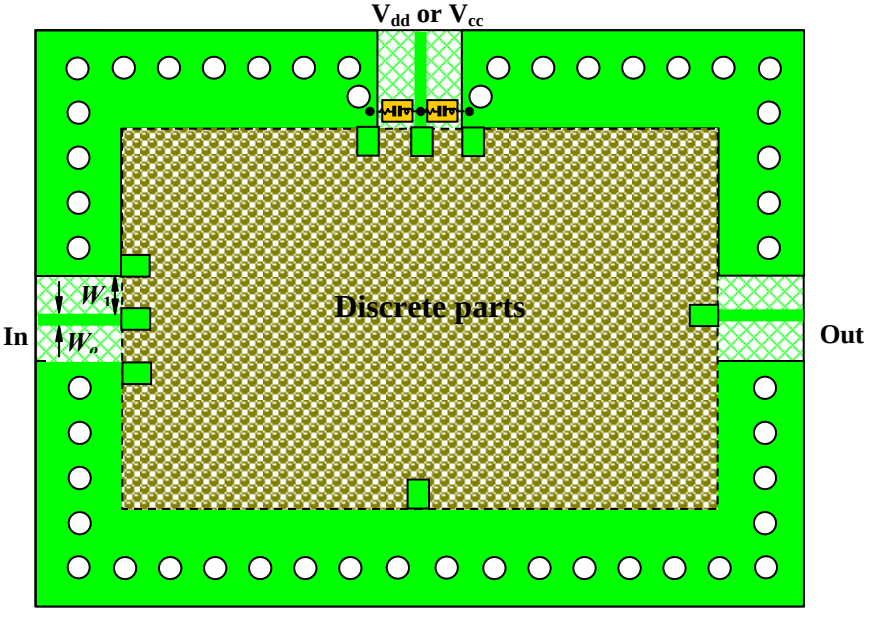


Figure 3.16 Layout of a tested PCB for an RF module built by discrete parts, with 50 Ω characteristic impedance of input and output runners.

Top metallic area
Bottom metallic area
Module built by discrete parts
50 Ω runner
Conductive via from top to bottom
"Zero" Capacitor

In Figures 3.13 to 3.16, only the main tested RF parts and the main ports, including input, output, and DC Power supply, are shown, while the circuit detailed items are neglected, since our discussion is confined on RF grounding.

In all PCBs as shown in Figure 3.13 to 3.16, the entire bottom side of the PCB serves as part of the grounding area. As for the top side of the PCB, it is preferred for the grounding area to have a metallic rectangular frame around the peripheral edges. In order to combine the grounding areas on the top and bottom sides, many conductive vias (holes) through the top and bottom sides are added. In experimental testing, the biggest difference of test results exists between two cases with and without conductive holes. For example, the insertion loss of a filter might be different from 1 to 2 dB between two cases with and without vias.

It is well-known that in an RF range, a conductive hole is equivalent to an inductor, a resistor, and a capacitor connected together in parallel. Their values mainly depend on the diameter of the holes and other parameters related to the PCB's material and configuration. The smaller the hole, the higher the equivalent inductance and the lower the equivalent resistance. It is therefore desirable to enlarge the hole as much as possible. Usually, the desirable diameter of a via, D, should be

$$D > 10 \text{ } mil\text{s}. \tag{3.32}$$

At the input, output, and DC power supply ports, one might like to paste more grounding area beside the runners. However, the spacing between the input or output runners and the adjacent grounded edge, W_1, must be wide enough so that the edge spray capacitance between the input or output runners and the grounded area in both sides can be neglected. For instance, as shown from Figure 3.13 to 3.16, the width of the input or output runners, W_o, must correspond to 50 Ω characteristic impedance of micro strip line because it is supposed to be connected to the input signal generator or output analyzer. In order to keep its characteristic impedance around 50 Ω after attaching more metallic area as the grounding area in its both sides, empirically the following condition must be satisfied:

$$W_1 > 3 \, W_o \, . \tag{3.33}$$

However, if the runner for the input and output is designed on the basis of a coplanar waveguide as shown in Figure 3.17, condition (3.33) is not necessary. As a result, more metallic area can be added to the grounded surface. Under the coplanar waveguide design, the width of input or output runners is usually narrower than that corresponding to a micro strip line with 50 Ω of characteristic impedance, W_o. Also, the spacing between the runner and the adjacent grounded edge is much narrower than $3W_o$. Consequently, the grounded area is expanded and has a simpler geometrical shape as shown as Figure 3.17.

In the case of COB testing as shown in Figure 3.13, one end of the runner is beveled and connected to the IC die by a bonding wire. The runner is bevel-shaped to avoid the jump of the characterized impedance along the transportation path. On the other hand, in order to reduce the parasitic capacitance between the runners and to have a smooth impedance change of the runner itself, it is preferred for the runners coming out from the IC die to form a "radiation" shape.

With the exception of the above effort for RF grounding, "zero" capacitors must be added between the V_{dd} or V_{cc} terminals and the adjacent grounding area because the V_{dd} or V_{cc} terminal must be AC or RF grounded as well. The grounding potential is supposed to be the same for all the operating frequencies. Should the circuit be involved in more than one operating frequencies, such as *RF*, *LO*, and *IF* frequency, more than one "zero" capacitor must be added correspondingly. When testing for the RFIC die, the "zero" capacitors should be put as near the IC die as possible.

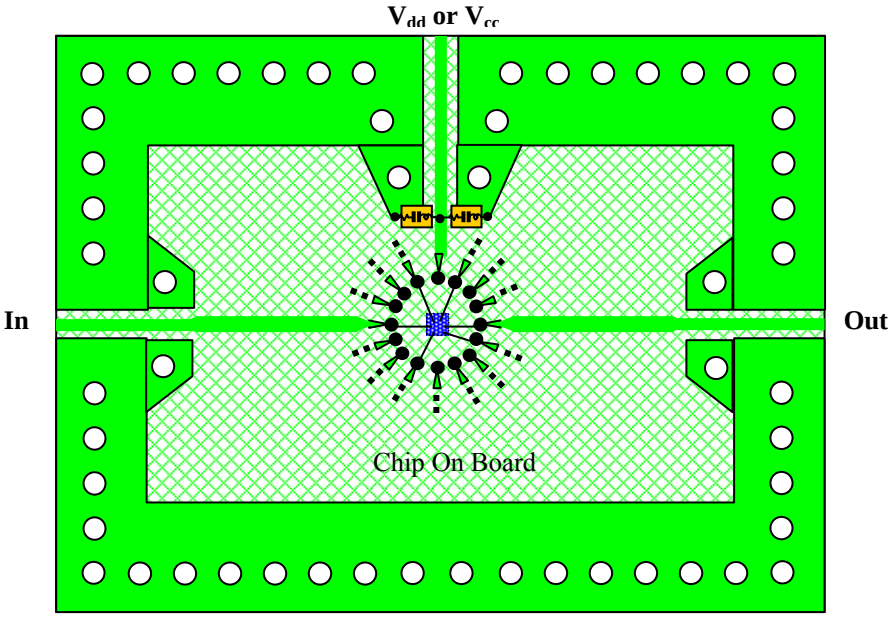


Figure 3.17 Layout of a tested PCB for RFIC COB testing with coplanar waveguide type of input and output runners.

Bond wire
Pad
IC die
Top metallic area
Bottom metallic area
50 Ω runner
End of a runner
Conductive via from top to bottom
"Zero" Capacitor

In addition to RF grounding, there are some other considerations for the improvement of a test PCB, such as, to have gold plating for the terminal runners, to add a resistance material on both top and bottom sides so as to protect the PCB from damage. However, these are beyond our topics of this section.

3.3.1.1.2 RF Grounding with a Rectangular Metallic Frame

All of the test PCB shown above have a rectangular metallic frame. It is emphasized as a special and important configuration of a test PCB. It functions not only for good RF grounding but partially also for shielding.

As shown in Figure 3.18, owing to the "zero" potential at the rectangular metallic grounded frame, most electric lines produced and then radiated either from the circuit inside the PCB or from any interference source outside the PCB would be terminated on the rectangular grounded frame. We say "most of electric lines" but not "total electric lines" because some of them would cross over the barrier of the rectangular grounded frame when the environment of the test PCB is somewhat complicated rather than a free space and, in addition, some of electric lines could cross over the frame through the plastic medium. Nevertheless, the isolation between the circuit on the PCB and outside world is greatly improved due to the existence of the rectangular grounded frame.

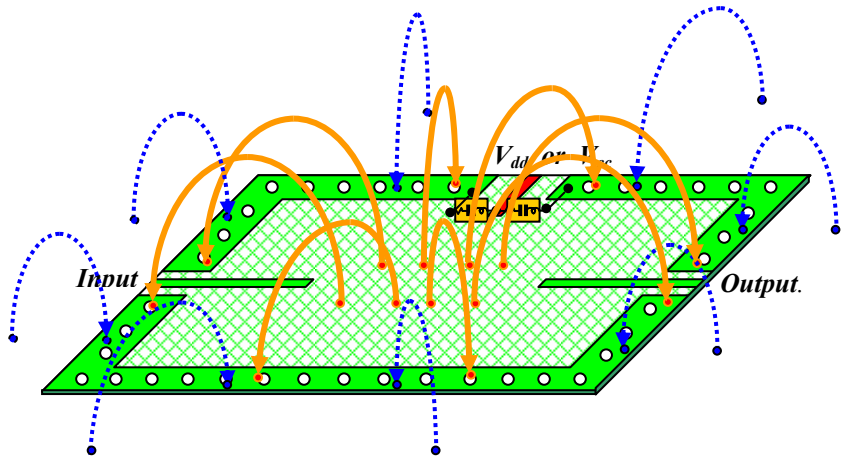


Figure 3.18 Electric lines radiated either from the test circuit inside the PCB or from any interference source outside the PCB would be terminated on the rectangular grounded frame.

3.3.1.1.3 An Example

Let's introduce an example of an imperfect RF grounding as shown in Figure 3.19. A test PCB of a transceiver is shown, on which an IC package with 20 pins is soldered on the FR4 PCB. Instead of detailed drawing, only those parts of layout related to the problems are depicted.

There are 3 main problems:

1) Poor RF grounding
   When the sensitivity test is conducted at the SMA terminal, it is found that the sensitivity can be varied from 2 to 6 dB if the point A shown in Figure 3.19 is additionally connected to either B or C by a short copper wire. Mostly it looks like that it results from the poor grounding for the SMA connector.

   Before the discussion of its grounding problem, it should be understood that a metallic area with high conductivity is usually not an equipotential surface in an

RF circuit block or system. Therefore, a special reference point must be assigned. In an RF circuit block or a system, the reference point of RF grounding for the entire test circuit board is usually assigned to the grounding point of the DC power supply and is denoted by "GND" in Figure 3.19.

From Figure 3.19, one can find that the metal portion denoted by "A" is useless in the connection of SMA connector to the "GND". The only way to connect the SMA ground to the "GND" is the path which can be described as

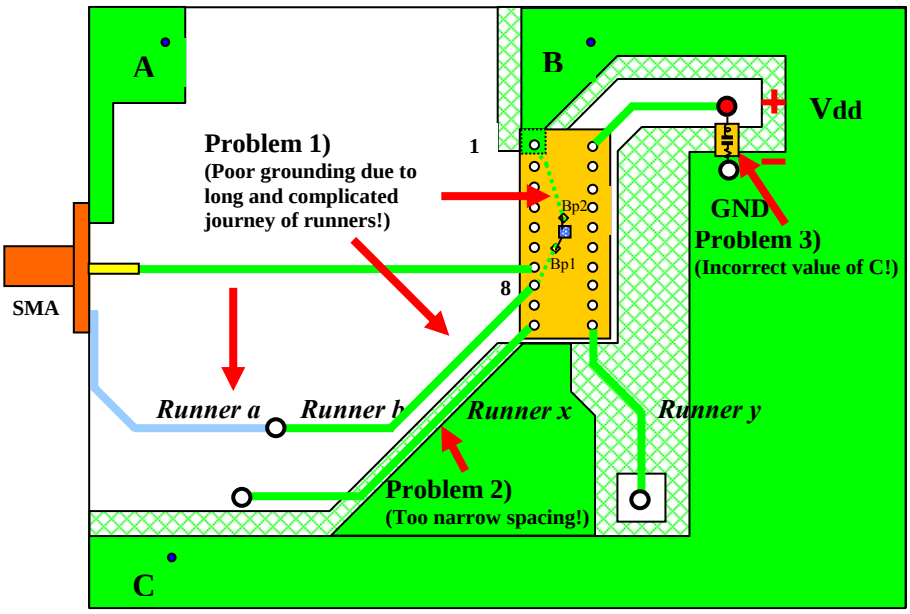


Figure 3.19 An example of test PCB problems in RF testing .

- Starting from the SMA connector, go through *runner a* on the bottom side to a conductive via from bottom to the top;
- From the via going along *runner b* on the top side all the way to pin#8 on the IC package;
- From pin#8, go through a tiny runner to the bond pad#1, *Bp1*, on the IC package;
- From *Bp1* on the IC package, go along a bond wire to the inlet bond pad on the IC die;
- On the IC die, go from inlet bond pad along the grounding path to the outlet bond pad;
- From outlet bonding pad on the IC die, go along a bond wire to another bonding pad on the package, *Bp2*;
- From pad#2, *Bp2*, go along a tiny runner to pin#1 of the IC package;
- From pin#1 go along the grounding path on the top metal portion denoted by B to the "GND," the grounding point of the DC power supply.

How long and complicated the journey is! All the segments of the path described above might contribute a considerable inductance in series and capacitance in parallel between the ground of the SMA connector and the ground of the DC Power supply, GND. These parasitic inductances and capacitances depend on the many parameters of PCB, IC package, vias, and all the runners. Obviously, the grounding of the SMA connector is poor!

2) Undesired parasitic capacitance
Undesired parasitic capacitance is found on the *runner x* as shown in Figure 3.19. This is due to the incorrect layout of the metallic grounding portion. The edge of *runner x* is adjacent to the grounding metal. The spacing between the *runner x* and the grounding metal is too little and brings about a fair amount of coplanar parasitic capacitance. The additional parasitic capacitance has been tested for *runners x* and *y*. The difference of the parasitic capacitances of these two runners is about 60 Ω at the carrier frequency of the transceiver. The *runner y* has less parasitic capacitance because the more spacing between *runner y* and the grounding metal.

3) Incorrect value of the "zero" capacitor
Finally, it is found that the value of the "zero "capacitor is incorrect. This brings about the leakage of the RF signal on the DC power supply wire.

The solution for the problems can be illustrated in Figure 3.20.

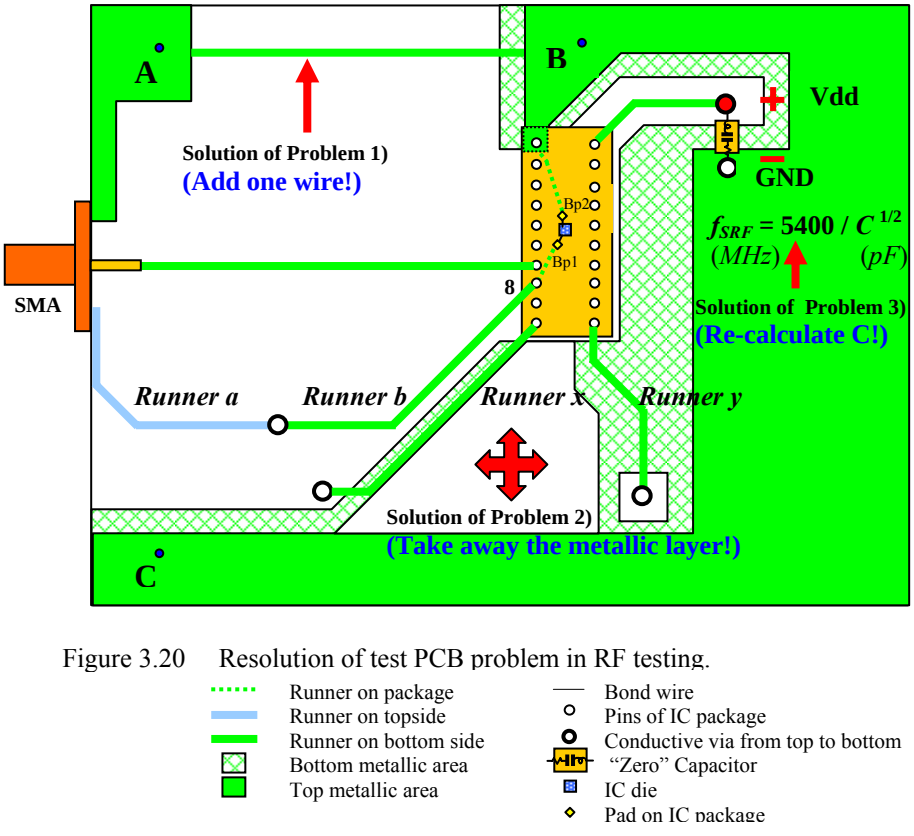


Figure 3.20 Resolution of test PCB problem in RF testing.

The solution for problem 1) is simply to make a connection between point A and point B. The test demonstrates that further improvement can be achieved if an additional connection between point A and C is made.

The solution for problem 2) is simple: as shown in Figure 3.20, take the metal portion adjacent the *runner x* away. Very often, some of new engineers try to fill as much grounding area as possible but forget the coplanar parasitic capacitance. A trade-off must be taken between these two aspects.

The solution for problem 3) is to carefully calculate the value of the "zero" capacitor on the basis of the discussion in chapter 4. It would be even better to apply "infinitive" inductors to assist the RF grounding as shown in Figure 3.3 if the "infinitive" inductors are available in your engineering stock room.

As a matter of fact, a test PCB as shown in Figure 3.21 is preferred, where the RF grounding consists of the bottom metal and a rectangular metallic frame on the top. The top and bottom metallic portions are connected together by many conductive vias. As mentioned above, the rectangular metallic frame plays not only for good grounding but also for better shielding. The test board with such a neat configuration has a better RF grounding and is widely adapted.

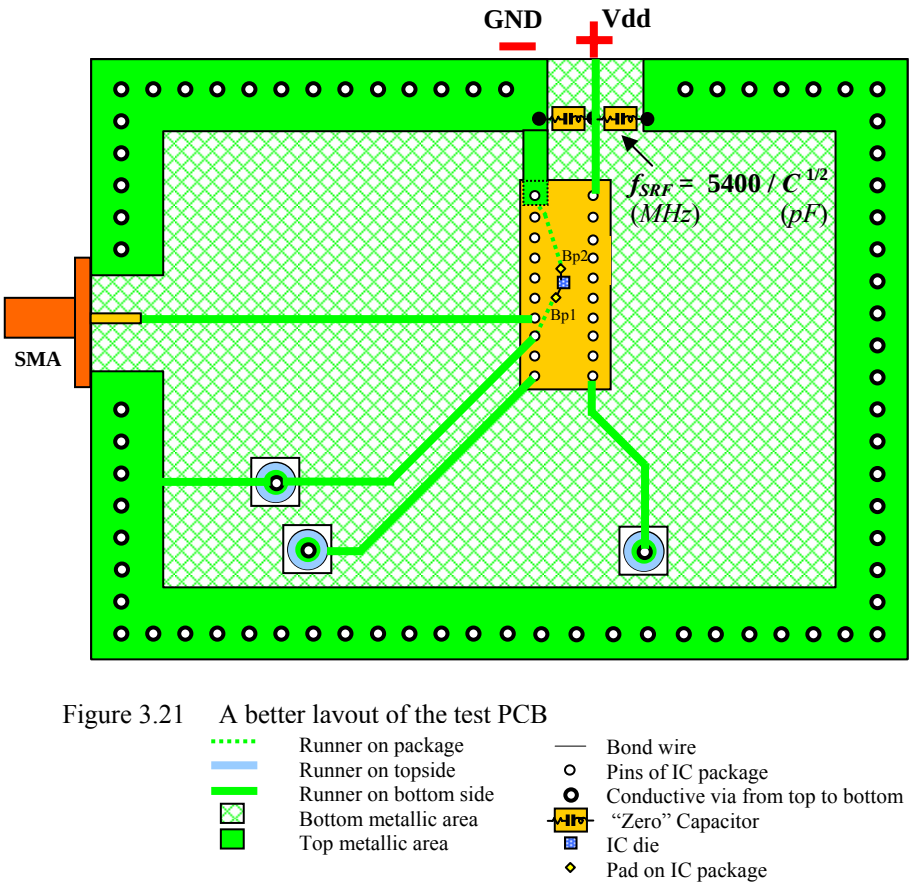


Figure 3.21 A better lavout of the test PCB

3.3.1.2 Large Test PCB

A test PCB with a large size usually contains various combinations of the 4 basic types of RFICs or RF modules.

The main difference of a test PCB due to its size is the potential of a metallic grounded surface, which needs to be an equipotential surface so as to function as a grounded area. At the least, the ports of input, output, DC power supply, and other control points where the test board is connected to other equipment must have their corresponding grounded ends simultaneously equipotent at "zero" potential. In an RF range, a metallic surface or plane made by material with high conductivity is an equipotential surface or area if the dimension of the test PCB is greatly lesser than the quarter wavelength of a 50 Ω runner. It is then a case under condition (3.30). However, it might not be an equipotential surface or area if the dimension of the test PCB is equal to or greater than the quarter wavelength of a 50 Ω runner. It is then a case under condition (3.31), which usually exists in the testing of a huge system with complicated circuitry. In concrete, we will focus on the discussion of how to force the entire rectangular metallic frame on the top side of the test PCB to be an equipotentially grounded surface, or the grounded ends at input, output, DC power supply, and other control points to be at an equipotentially grounded level.

Figures 3.22 to 3.26 illustrate the main points in the design of a large PCB. They show how to maintain the entire rectangular metallic frame on the top side of the test PCB to an equipotential ground surface by using "zero" capacitors, half-wavelength runners, half-wavelength cables, quarter-wavelength runners and quarter-wavelength cables, respectively.

Similar to Figures from 3.13 to 3.16, only the main tested RF part and the main ports, including input, output, and DC Power supply, are shown in Figures 3.22 to 3.26, while the circuit detailed items are neglected, since our discussion is confined on RF grounding.

3.3.1.2.1 RF Grounding by "Zero" Chip Capacitors

When the dimensions of the test PCB are comparable or larger than the quarter wavelength at the operating frequency, RF grounding can be accomplished by means of "zero" capacitors as illustrated in Figure 3.22. The metallic grounding portion is cut into many small pieces, which must be much smaller than the electric quarter wavelength. In terms of such a scheme all the terminals, including V_{dd}, input and output, have almost same RF reference electric potential at the operating frequency. This technology seems to be easier and more convenient than the addition of runners or cables with quarter- or half-wavelengths, though many "zero" capacitors must be applied.

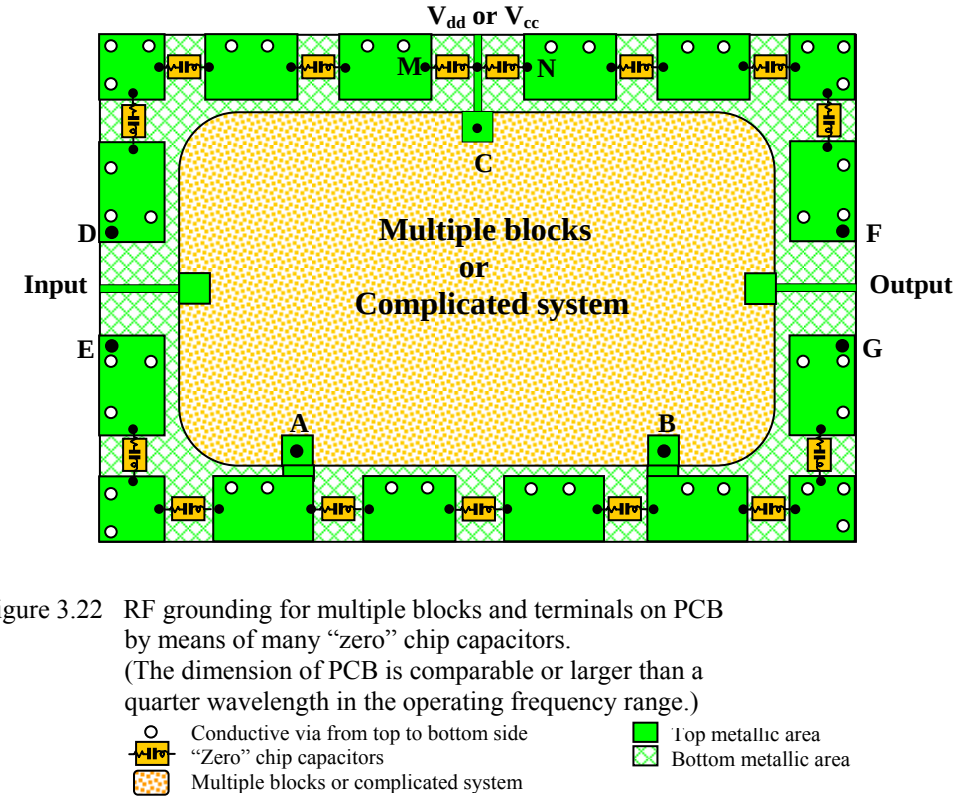


Figure 3.22 RF grounding for multiple blocks and terminals on PCB by means of many "zero" chip capacitors. (The dimension of PCB is comparable or larger than a quarter wavelength in the operating frequency range.)

3.3.1.2.2 RF Grounding by a Runner or a Cable with Half or Quarter Wavelength

Similar to the RF grounding scheme as shown in Figure 3.22, the goal of RF grounding is to enable the points, A, B, C… F, G, M, N, to be equipotentially grounded points. Instead of applying "zero" capacitors, however, the half-wavelength runners or cables as shown in Figures 3.23 and 3.24 respectively, can force these points to be at an equipotentially grounded level because the voltage or electric potential is the same at both ends of a half-wavelength runner or cable. Two "zero" capacitors are connected between V_{dd} and point M and N. Thus, point *M* and *N* can be accounted as primary grounding points.

A half-wavelength runner cable is somewhat lengthy. An alternative would be to replace them by quarter-wavelength runners and cables as shown in Figures (3.25) and (3.26), respectively. One end of the quarter-wavelength runner or cable is opened while another end is connected whichever point needs to be grounded. Theoretically it will force the connected end to be grounded because the other end is open, and thus, its impedance is infinitive.

In the various RF grounding of a test PCB with large sizes, the most popular technology is to use "zero" chip capacitors because of its relatively small size compared to that of a half/quarter-wavelength runner or cable. The half/quarter-wavelength cable is too clumsy and thus is seldom applied in practical engineering.

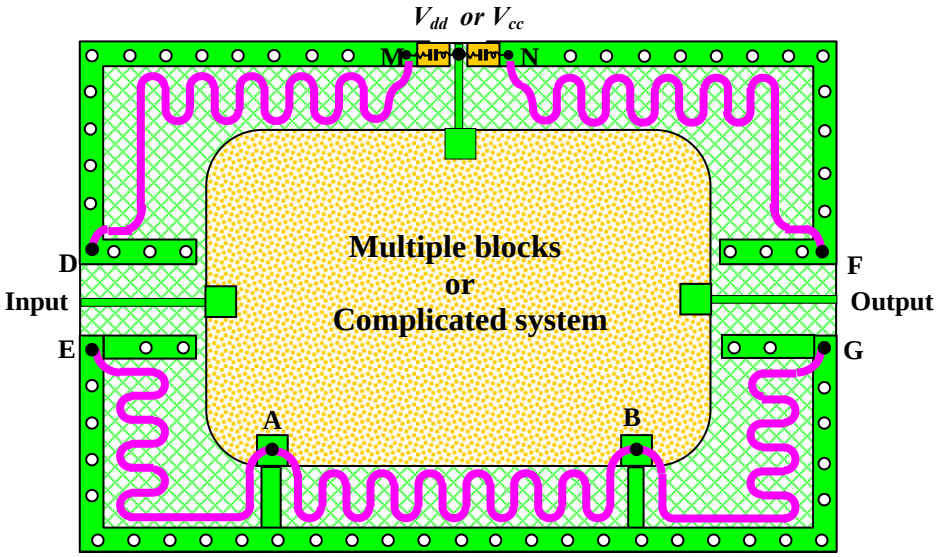


Figure 3.23 RF grounding for multiple blocks and terminals on a PCB simultaneously by means of a half-wavelength runner.
(The dimension of PCB is comparable or larger than a quarter-wavelength in the operating frequency range.)

Length of runner = λ/2
Conductive via from top to bottom side
Multiple blocks or complicated system
Top metallic area
Bottom metallic area
"Zero" capacitor

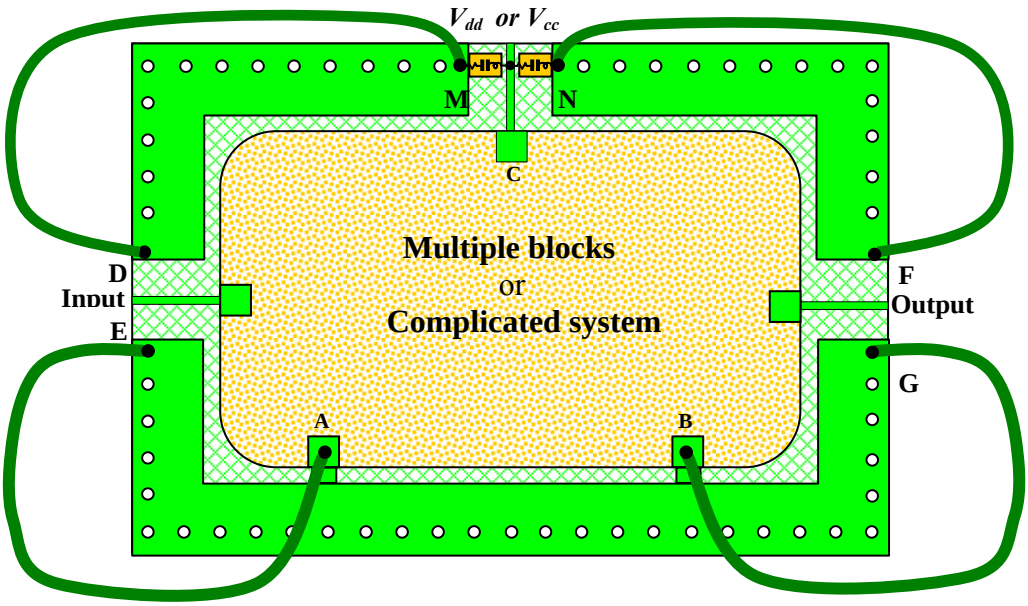


Figure 3.24 RF grounding for multiple blocks and terminals on a PCB simultaneously by means of a half-wavelength cable.
(The dimension of the PCB is comparable or larger than a quarter-wavelength in the operating frequency range.)

Length of cable = λ/2
Conductive via from top to bottom side
Multiple blocks or complicated system
Top metallic area
Bottom metallic area
"Zero" capacitor

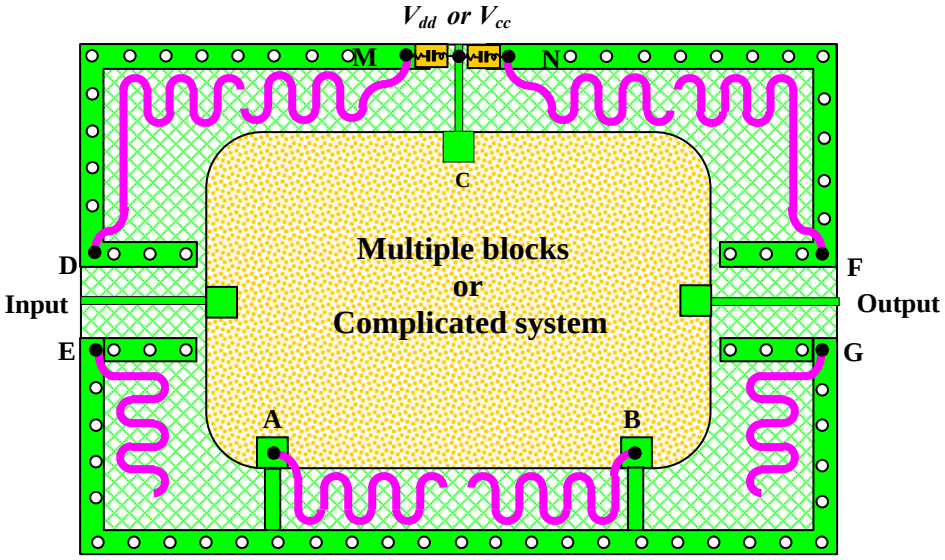
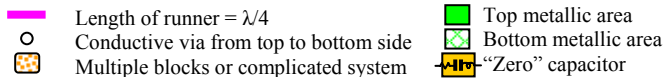


Figure 3.25 RF grounding for multiple blocks and terminals on a PCB simultaneously by means of a quarter-wavelength runner.
(The dimension of PCB is comparable or larger than a quarter-wavelength in the operating frequency range.)



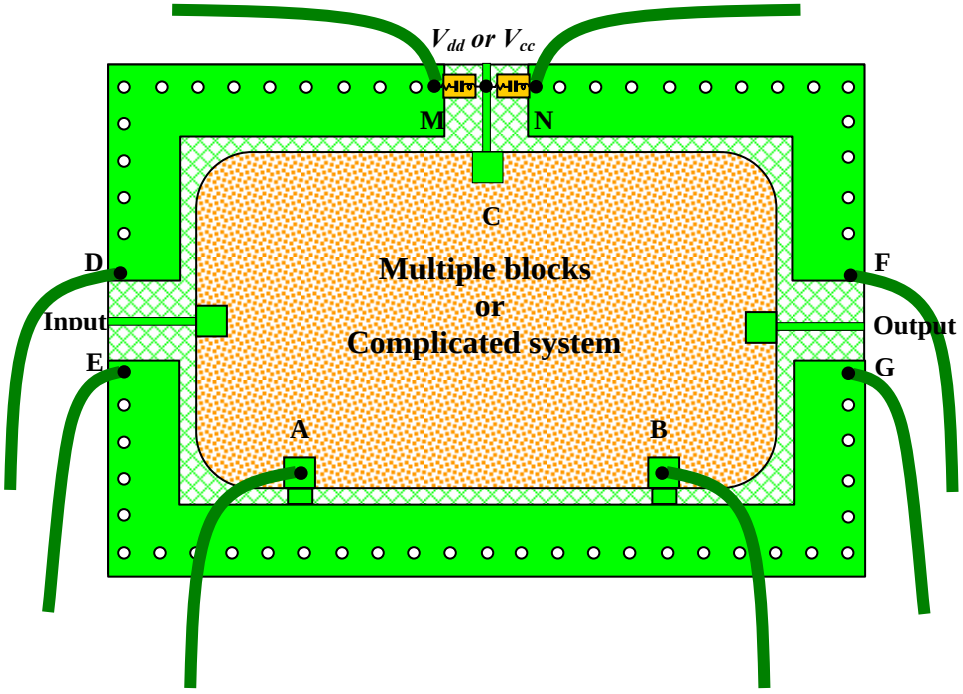
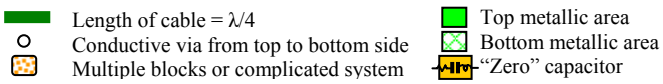


Figure 3.26 RF grounding for multiple blocks and terminals on a PCB by means of a quarter-wavelength cable.
(The dimension of PCB is comparable or larger than a quarter-wavelength in the operating frequency range.)

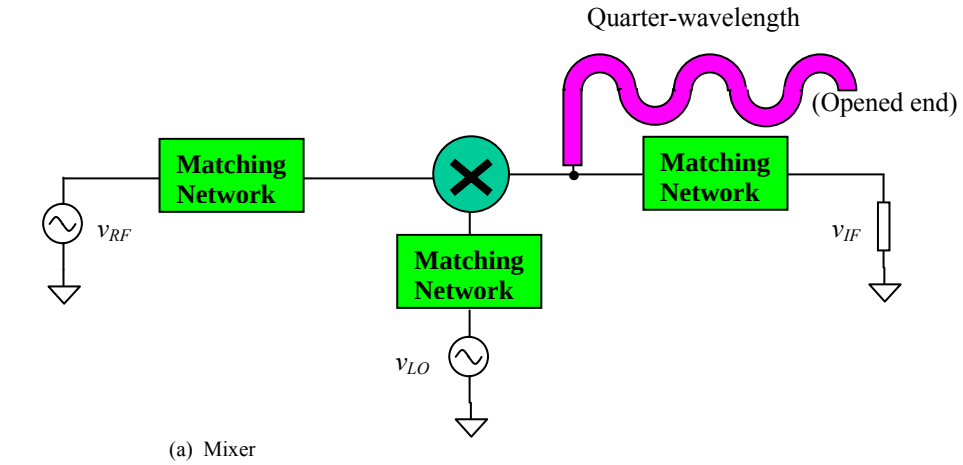


3.3.2 Isolation between Input and Output in a Mixer or an Up-converter

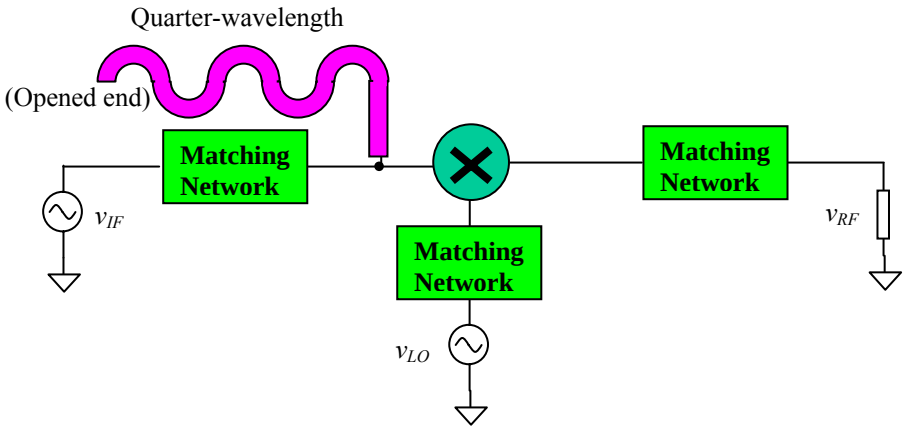
In either mixer or up-converter designs, the isolation between input and output is one of the important parameters to be faced with. The LO injection at the LO port could leak to the RF or IF ports. Similarly, the RF power from RF port could leak to the RF or IF ports.

The LO power leakage into the IF portion dominates over the RF power leakage to the IF port because the LO injection has the highest power among these 3 ports.

By means of a quarter-wavelength micro strip line, a mixer can perform with excellent isolation between the LO, RF and IF ports. Figures 3.27 (a) and (b) show the schemes in mixer and up-converter designs, respectively.



(a) Mixer



(b) Up - converter

Figure 3.27 Isolation schemes using quarter-wavelength micro strip lines in a mixer or an up-converter

A quarter-wavelength micro strip line corresponding to the LO frequency is connected to the IF port. Its unconnected end is in an opened-circuit state. Consequently, the IF port is forced into zero voltage at the connected point at the LO

frequency. The frequency difference between the LO and RF ports are equal to the IF frequency. The micro strip line with a LO quarter-wavelength can usually play the same role for the isolation between the RF and IF ports at the same time, as long as its *Q* value is not too high. In general, the isolation can be enhanced by more than 20 *dB* with this technology.

3.3.3 Calibration for Network Analyzer

The network analyzer is a powerful tool in the measurement of a RF circuit block or system; calibration is an important step before the test or measurement is conducted.

One can carry out the calibration by means of the standard calibration kit, which is provided by the manufacturer. There are 4 basic calibration procedures. They are "open", "short", "50 *Ω*", and "through". However, more procedures may be conducted for a higher precision in measurement.

Instead of the standard calibration kit provided by the manufacturer, a more precise way of measuring is to develop the calibration kit by the RF design engineer as shown in Figure 3.28.

Figure 3.28 (a) and (b) depict the layout for a DUT, in which the DUT is located at the center of the board and there is a micro strip line leading from the input and output port to the circuit block. The characterized impedance of the micro strip line is 50 *Ω*. With the same PCB as adapted for the DUT, a calibration kit is developed: its layout is shown in Figure 3.28 (c) and (d). It consists of 4 sub-calibration kits for the "open", "short", "50 *Ω*", and "through" calibration purposes. Each sub-calibration kit has the same grounding pattern and size as it does for the DUT. By this way the calibrations are closer to the actual environment of the DUT than the standard calibration kit does.

It should be noted that in order to ensure the resistors of 50 *Ω* unchanged within a desired frequency bandwidth, a combination of resistors might be necessary. For instance, a 50 *Ω* resistor could be substituted by two 50 *Ω* resistors in series with another 100 *Ω* resistor in parallel. This combination of multiple resistors tends to cancel the frequency variation of the resistance in the resulting resistor. Empirically the value of 50 *Ω* could be kept unchanged from DC up to 6 *GHz*.

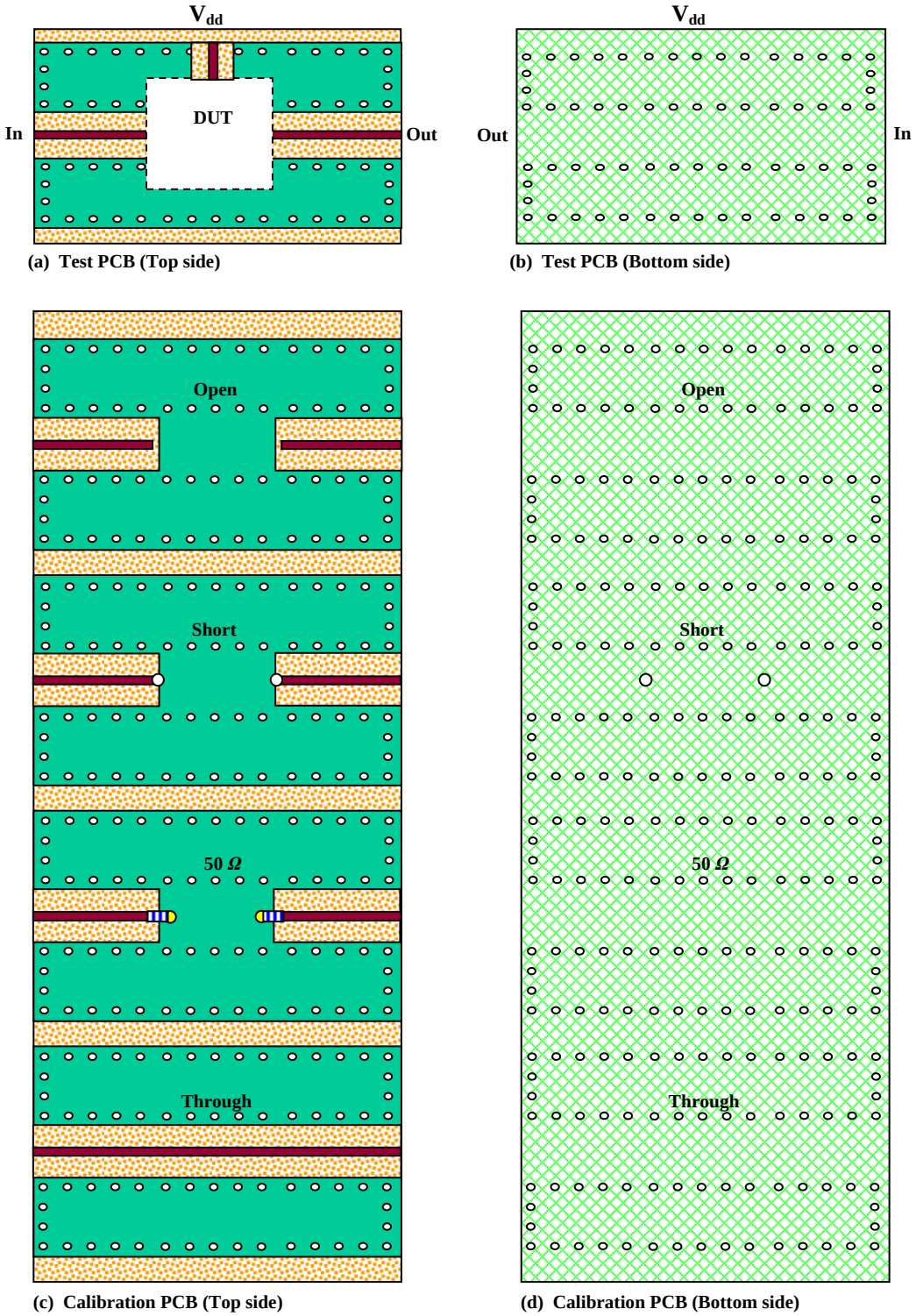


Figure 3.28 Layout of PCB for Network Analyzer calibration

PCB
Top metallic area
Bottom metal area
Copper with gold plating
Soldering point
Conductive hole from top to

3.4 RF Grounding for Reduction of Return Current Coupling

When circuitry is discussed, much attention is paid on the forward current that the circuit draws from the DC power supply. Very often, however, the return currents, which originate from the grounded points of the circuit to the grounded point of the DC power supply, are ignored. In an actual RF layout or a test board, the return current could completely screw up the performance of the circuits due to the coupling between blocks through the return current. It seems to be a dark corner especially for a new engineer who has little experience in the layout.

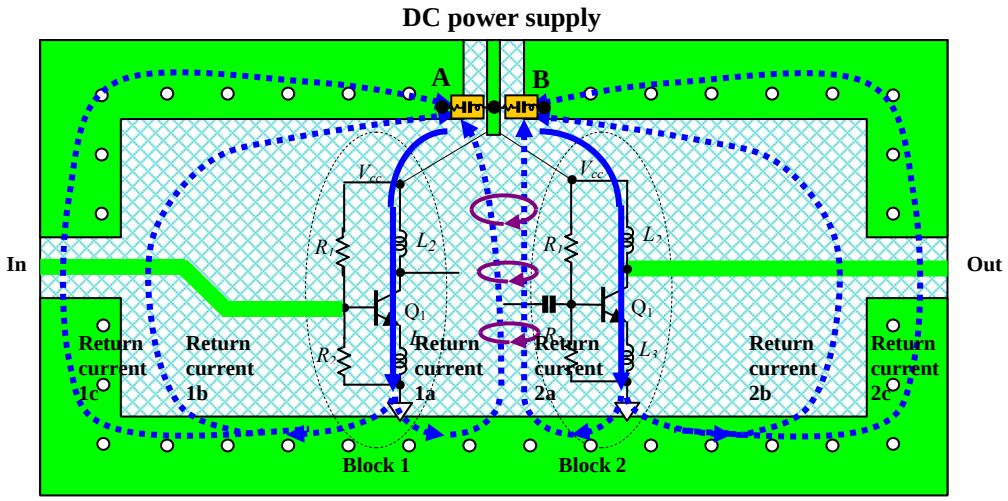
3.4.1 A Circuit Built by Discrete Parts on a PCB

Figure 3.29 (a) depicts a circuit consisting of two blocks, built by discrete parts on a PCB. Its bottom side is entirely covered with copper and connected to the top side's metallic (copper) portions by many conductive vias, so that all the copper portions together are supposed to be a good DC grounded surface. By means of the "zero" capacitors connected between the DC power supply and its adjacent grounding points, A and B, there is no build-up of RF voltage between the DC power supply and the grounded points, A or B. Furthermore, there is no build-up RF voltage in the entire metallic portion as long as the dimension of the PCB is much shorter than the electric quarter wavelength. Rather than the existence of coupling between these two blocks, the performance of the circuit is expectedly perfect.

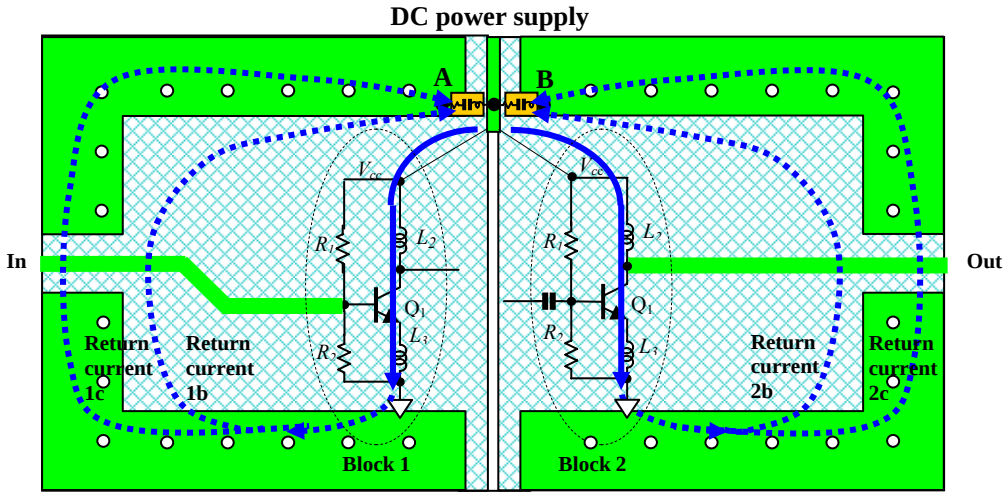
In reality, the currents flowing on the PCB form a complicated pattern, depending on the placement of parts and the arrangement of the grounding areas. All currents drawn from the DC power supply must be returned to their adjacent grounding points, A or B. In Figure 3.29 (a), the solid arrow lines approximately represent the current drawn from the DC power supply while the dashed lines represent the return currents, which flow from the grounding points of the blocks to the grounding points of DC power supply, A and B. Presumably, the return current can be demarcated into three branches in each block, that is, 1*a*, 1*b*, 1*c* and 2*a*, 2*b*, 2*c*. The currents 1*a* and 2*a* are the currents returning from block 1 and 2's grounding points and flowing through the middle portion of the bottom side copper together. The currents 1*b* and 2*b* are the currents returning from block 1 and 2's grounding points and flowing through the left and right portion of the bottom side copper respectively. The currents, 1*c* and 2*c*, are the currents returning from block 1 and 2's grounding points and flowing through the left and right portion of the top side copper respectively. Obviously, the coupling or cross-talk between the return currents 1*b*, 1*c*, and 2*b*, 2*c*, might be negligible as long as the distance between these two group paths is far enough whereas the coupling or cross-talk between the return currents 1*a* and 2*a* might be very strong since they are flowing together.

The cross-talk between 1*a* and 2*a* is magnetic coupling. Their interaction is equivalent to an appreciable feedback from block 2 to block 1; this eventually degrades the performance of the circuitry. A simple way to reduce such coupling is to cut a slot in the middle copper portion on the bottom side of the PCB so as to eliminate or greatly

reduce the return currents 1*a* and 2*a*, as much as possible. Figure 3.29 (b) presents such an idea.



(a) Incorrect layout



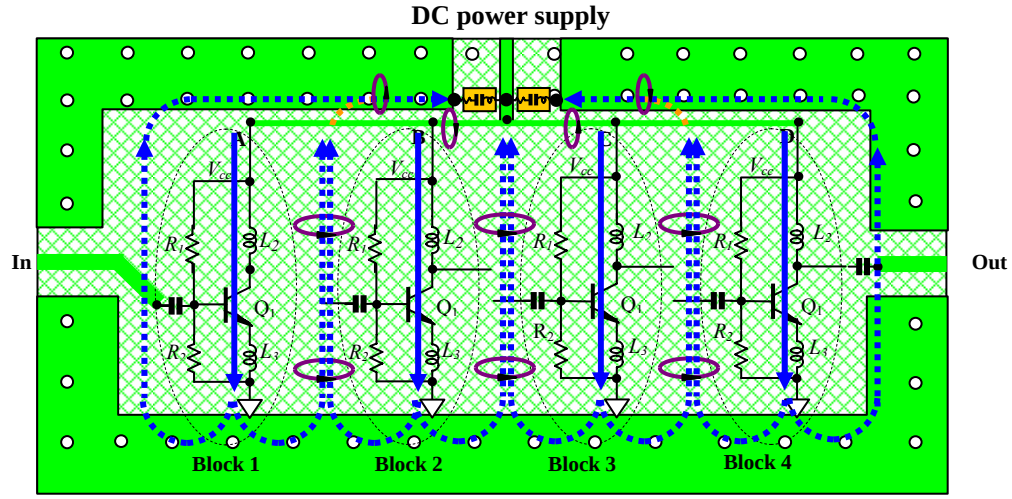
(b) Improved layout

Figure 3.29 Return current and grounding of two block RF circuit with discrete parts

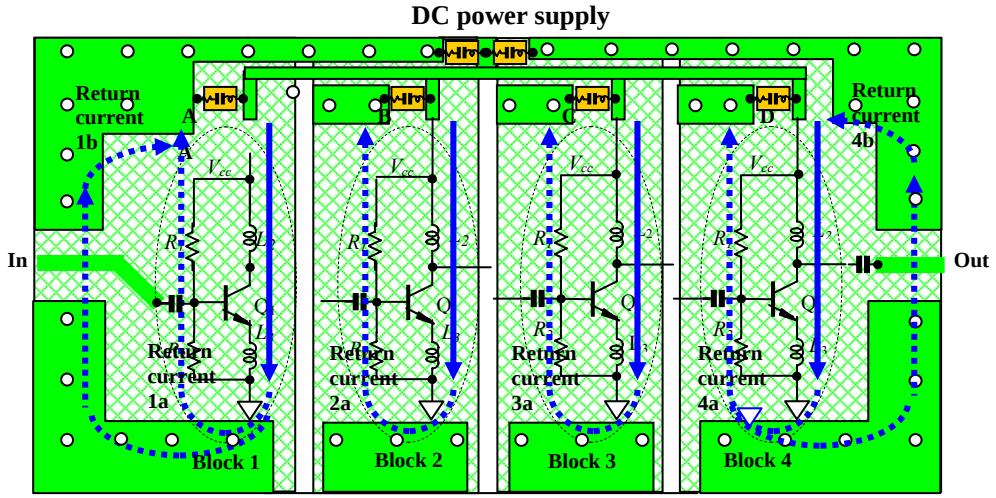
"Zero" capacitor
Forward current
Return current
Coupling of flux
Top metallic area
Bottom metallic area
Conductive via from top to bottom side

When a circuit consists of more than two blocks which are built by discrete parts on a PCB as shown in Figure 3.30(a), the "cut-off" slot must be repeated between every two adjacent blocks as shown in Figure 3.30(b) in order to eliminate or reduce the coupling or cross-talk from the return current.

On the other hand, attention should be paid to the forward current coupling or cross-talk due to the common DC power supply as shown in Figure 3.30(a). A possible solution is to provide the DC power supply to each block separately on the PCB and then add many "zero" capacitors so as to guarantee the impedance between each DC power supply for each block to their adjacent grounding points A, B, C, and D, respectively fairly close to zero. The metallic area between "zero" capacitors must be small enough so that all the small metallic pieces between the "zero' capacitors are approximately in the same zero electric potential.

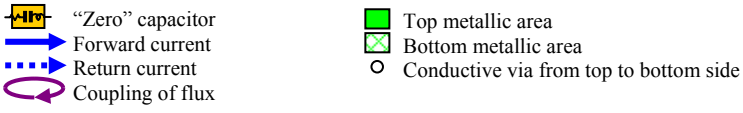


(a) Incorrect layout



(b) Improved layout

Figure 3.30 Return current and grounding of RF multiple block circuit with discrete parts



3.4.2 RFICs

Now let's move on to the layout for RFICs.

In a simpler case where a circuit consists of only one or two blocks, the layout for the reduction of return current coupling could be similar to that of a circuit built by discrete parts as introduced above. In complicated cases, the layout must be carefully treated. Let's introduce a specific and somewhat complicated example as shown in Figure 3.31 to 3.34, a circuit containing 4 differential stages.

For simplicity, only the layout portions related to RF grounding and the return currents are shown in Figure 3.31 to 3.34 while the other portions are replaced by the schematic of their topologies.

From the viewpoint of symmetry, some designers apply a "highway" type of RF grounding as shown in Figure 3.31. The name "highway" is derived from the fact that the grounding portion is a long runner all the way from the right to the left. This grounding "highway" demarcates the whole IC die into two equal portions: the upper portion and the bottom portion. Two differential branches are symmetrically built in these two portions respectively. This type of layout keeps the geometrical symmetry to the differential pairs very well. However, it brings about the magnetic coupling in the "highway" between the return currents from different differential pair. From Figure 3.31 it can be seen that more and more return currents from individual differential pair flow into the "highway" as the resultant return current flows more close to the grounding point of the DC power supply. Consequently the strongest magnetic coupling is happened around the grounding point of the DC power supply, where the return current from all the individual differential pairs are coupled from

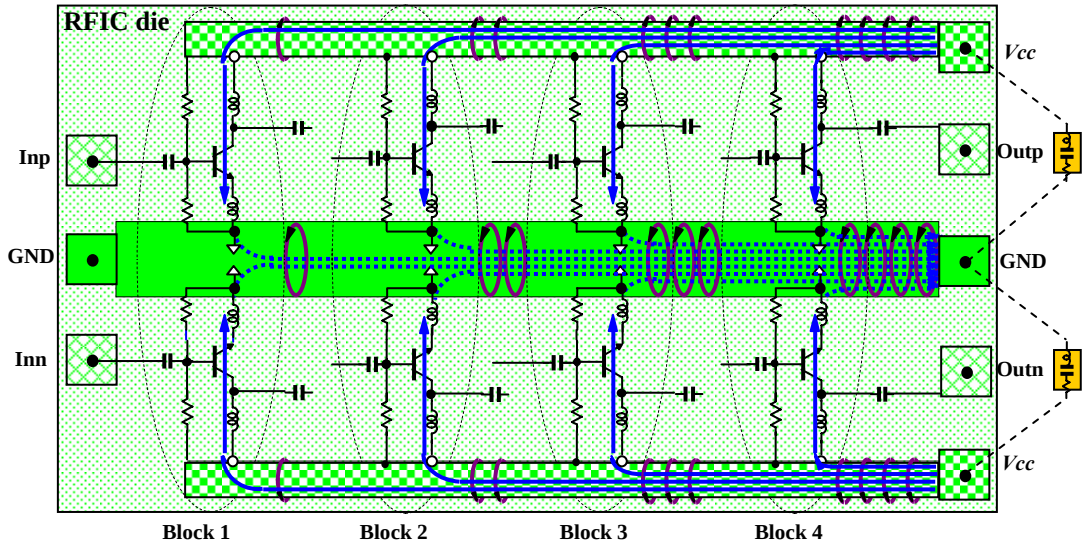
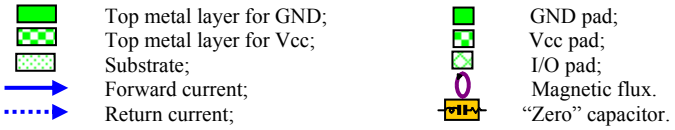


Figure 3.31 Forward and return current in a RFIC with "Highway" type of RF grounding.



each other. The magnetic coupling of return currents between different differential pairs could significantly impact the performance of the entire circuitry. Similarly, the magnetic coupling between the forward currents to the different differential pairs is also existed. From Figure 3.31 it can be seen that the "Highway" for the forward currents is just half of "Highway" for the return currents.

The layout with "Highway" configuration as shown in Figure 3.31 might be somewhat improved by the layout as shown in Figure 3.32. First, instead of the "Highway" grounding, the grounding of each stage to the grounding point of DC power supply is connected by an individual runner so that the return current would not flow on the common "Highway" but on the individual "Subway" where the coupling is reduced. The same scheme is applied to the forward current paths so as to reduce the forward current coupling. Secondly, in order to further reduce the coupling of either return or forward currents, the individual runners are simply kept not in parallel from each other as possible though they are running in the same direction. Of course, the price of the lower coupling is that more IC die area must be provided.

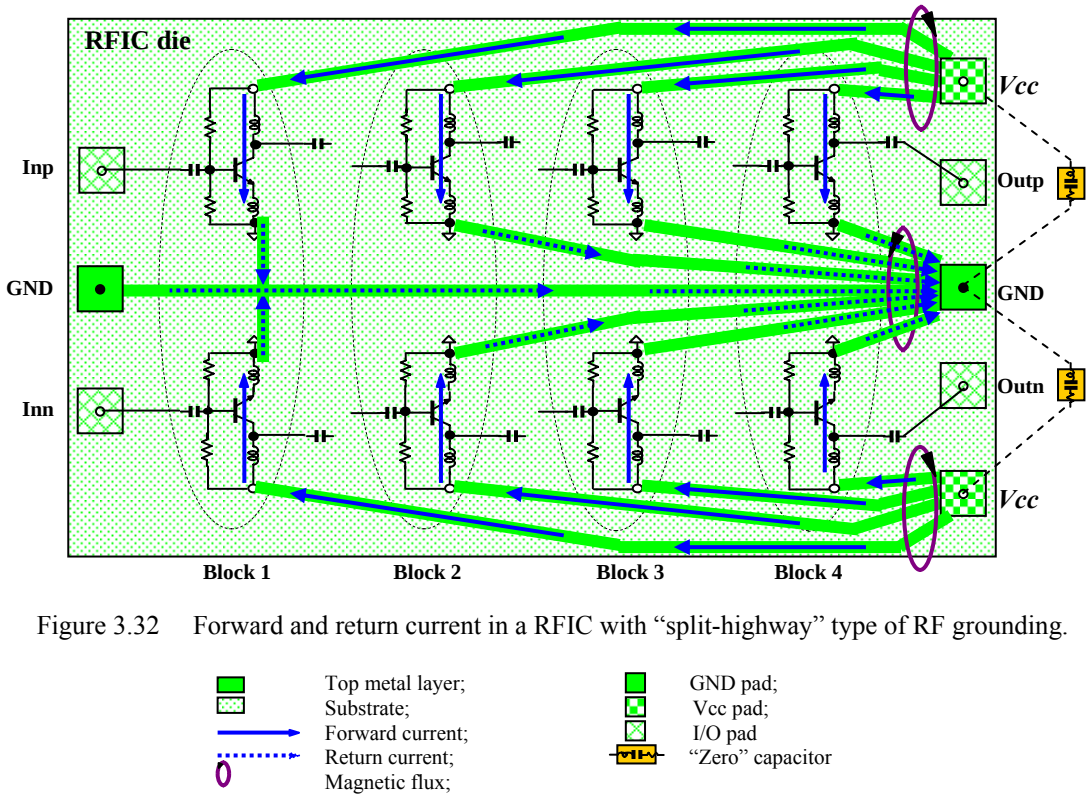


Figure 3.32 Forward and return current in a RFIC with "split-highway" type of RF grounding.

Another type of RF grounding for circuit containing 4 differential stages is shown in Figure 3.33. Its main feature is that there are two rectangular metallic frames around the periphery of the IC die and are overlapped all the way. One of them is the frame for RF grounding and another one is the frame for DC power supply. Instead of worrying about the coupling between either return or forward currents, the primary clue of the designer is to build a distributed capacitance between the RF grounding terminals and the DC power supply terminals by these two overlapped metallic frames, so as to bypass the AC or RF signals from V_{cc} or V_{dd} to the ground all the way. Also,

it is expected that the reduction of magnetic coupling since the return and forward current are overlapped together but their directions are exactly the reverse.

The performance of this overlapped type of RF grounding might be far from the primary expectations if the total length of the frame is about or greater than the quarter wavelength electrically. In this case the two overlapped metallic frames are a complicated system with micro strip line configuration. It creates not only the distributed capacitance but also distributed inductance. The metallic frame is not an equipotential surface if the distributed inductance is conceivable. On the other hand, the magnetic coupling between the either return or forward currents from different differential pair is not improved from that in the "highway" type of RF grounding. As shown in Figure 3.33, it just simply divide the coupling due to return currents in the "highway" into two halves and put them onto the upper and bottom side of the frame respectively. In either upper or bottom side of the frame, the magnetic coupling becomes more complicated by the mixing of return and forward currents together, since the upper and bottom branch of each differential pair are interlaced from each other in the connection for their return current and forward current to two metallic frame.

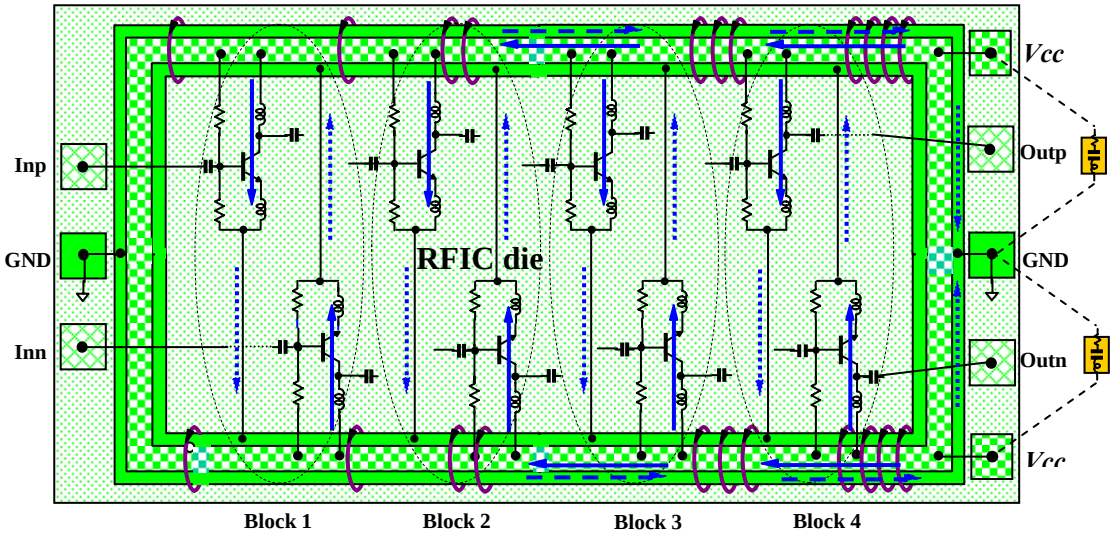


Figure 3.33 Forward and return current in a RFIC with "overlapped-ring" type of RF grounding.

The better RF grounding for reduction of either return or forward current coupling is to apply more grounding and DC power supply pads on the RFIC die. Figure 3.34 illustrates the idea for its layout. The purpose by using more pads is to avoid either return current coupling or forward current coupling between differential pairs. There are two grounding pads for return currents: one for the two differential pairs in the left side and another for the two differential pairs in the right side. The return current is individually returned to the grounding pad from each differential pair. There are two Vcc pads: one for the upper branches of the differential pairs and another for the bottom branches of the differential pairs. The forward current is individually delivered

from the Vcc pad to each branch of the differential pair. Consequently, the magnetic coupling in either return or forward currents can be reduced down to a negligible level.

In order to ensure the AC or RF equipotential at all the pads, 8 “zero” capacitors outside the RFIC die must be applied to the pads as shown in Figure 3.34. There are another 4 grounding pad being added for the testing of the RFIC, from which the probes of the IC probing station senses the IC die for the test equipments. The “zero” capacitors can be replaced by a half wavelength of micro strip lines or cables though the replacement is reluctant. The big size of a half wavelength of micro strip line or cable is usually unacceptable.

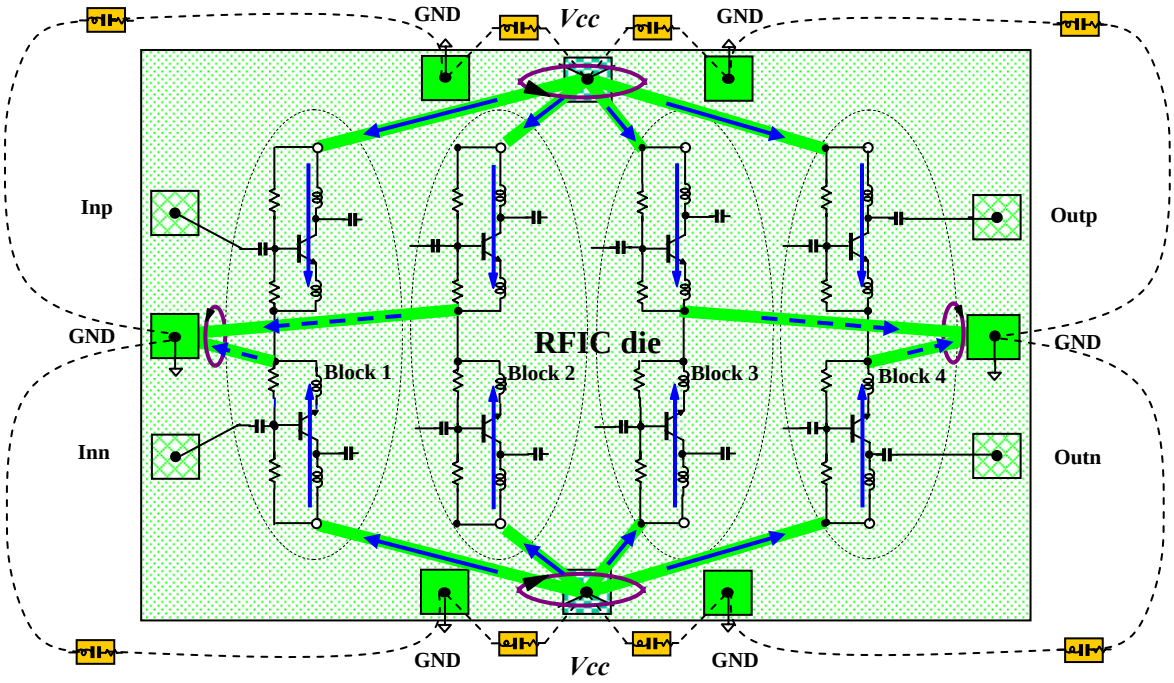


Figure 3.34 Forward and return current in a RFIC with “separated DC supply” type of RF grounding.

References

[1] Edward C. Jordan, “Electromagnetic Waves and Radiating Systems,” (Book), Second Edition, Prentice-Hall Inc., 1968.

[2] Ralph S. Carson, “High-Frequency Amplifiers,” (Book), , John Wiley & Sons, Inc., 1975.

[3] Herbert Taub and Donald Schilling, “Digital Integrated Electronics,” (Book), McGraw-Hill, Inc., 1977.

[4] Floyd M. Gardner, “Phaselock Techniques,” (Book), John Wiley & Sons, Inc., 1979.

[5] Carver Mead and Lynn Conway, “Introduction to VLSI Systems,” (Book), Addison-Wesley Publishing Company, 1980.

[6] Guillermo Gonzalez, “Microwave Transistor Amplifiers,” (Book), Prentice-Hall Inc., 1984.

[7] Paul R. Gray, Paul J. Hurst, Stephen H. Lewis, and Robert G. Meyer, “Analysis and Design of Analog Integrated Circuits,” (Book), Fourth Edition, John Wiley & Sons, Inc., 2001.

[8] Thomas H. Lee, “The Design of CMOS Radio-Frequency Integrated Circuits,” (Book), Cambridge University Press, 1998.

[9] Hannu Tenhunen, “CMOS Interconnects,” (Lecture), Electronic System Laboratory, Kungl Tekniska Hogskolan, 2000.

[10] Shim H.W., Zeeff, T.M., and Hubing T.H., “Decoupling strategies for printed circuit boards without power planes,” 2002 IEEE International Symposium on Electromagnetic Compatibility, , Vol. 1 , pp. 258-261, August 19-23, 2002.

Index

Contents

Chapter 4 Equivalent Circuits of Passive Chip Parts

Three passive chip parts to be discussed in this chapter are chip capacitor, chip inductor, and chip resistor. They are the main discrete parts and are widely applied in the implementation of RF modules.

It is a general trend to replace a discrete module by an IC (integrated circuit) and consequently, the market of passive chip parts would be gradually shrunk since the development of IC technologies in the past decades. However, the passive chip parts still play important roles in the electronic products since the discrete modules are still applied in the circuit implementation nowadays. On the other hand, by means of their special characteristics of self resonance, the chip capacitor and the chip inductor play important role for the AC or RF grounding no matter an electronic product is built by discrete parts or by an IC technology. Study of the passive chip parts is important to the designers of RF modules and RFICs. Especially, it is important in the test PCB design.

As mentioned in chapter 3, RF grounding, a “zero” capacitor or an “infinitive” inductor is generally more favored than a micro strip line or an RF cable. As a matter of fact, either “zero” capacitors or “infinitive” inductors are selected from chip capacitors and chip inductors, respectively. This is why this section is specially arranged.

We are intentionally going to expand our objectives of study to cover not only chip capacitors and chip inductors, but also chip resistors for other important purposes in RF design.

4.1 Modeling of Chip Parts

There are many different models for discrete passive parts. Some conventional models of a capacitor, inductor, and resistor are somehow complicated. In general, a complicated passive model, which is implemented by capacitors, inductors, and resistors, could be simplified to a simplest network, which consists of only one capacitor, one inductor, and one resistor.

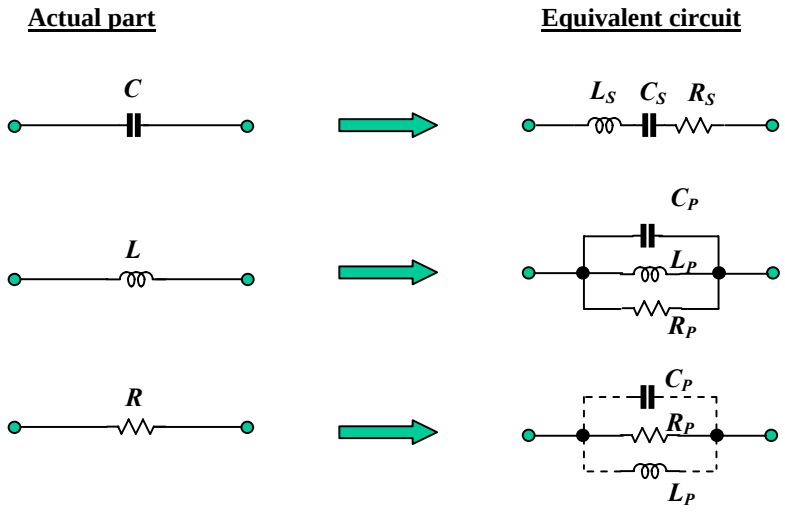


Figure 4.1 Models of chip capacitor *C*, chip inductor *L*, and chip resistor *R*.

Figure 4.1 shows the models we adapt for chip capacitors, chip inductors, and chip resistors.

A chip capacitor is represented by a series circuit of an ideal capacitor C_s, an ideal inductor L_s, and an ideal resistor R_s. The capacitor C_s, is an ideal capacitor with the capacitance specified by the manufacturer. The inductor L_s, represents the induced inductance of the AC or RF performance. The resistor R_s, represents the loss of power due to the current leakage, medium absorption, or the edge effect. The equivalent circuit of a chip capacitor has a self-resonant frequency since it consists of not only a capacitor but also an inductor. Its impedance approaches to zero at the resonant frequency since it is an *LC* resonant circuit in series. The self-resonance exists practically. When the operating frequency is lower than its self-resonant frequency, a chip capacitor is basically featured as a capacitor.

A chip inductor is represented by a parallel circuit of an ideal inductor L_p, an ideal capacitor C_s, and an ideal resistor R_s. The inductor L_p, is an ideal inductor with the inductance specified by the manufacturer. The capacitor C_p, represents the spray capacitance between the windings. The resistor R_p, represents the loss of power due to the current leakage, medium absorption, or the edge effect. The equivalent circuit of a

chip inductor has a self-resonant frequency since it consists of not only an inductor but also a capacitor. Its impedance approaches infinity since it is an *LC* resonant circuit in parallel at the resonant frequency. The equivalent resonant circuit in parallel reflects this character of infinitive impedance, which exists practically. When the operating frequency is lower than its self-resonant frequency, a chip inductor is basically featured as an inductor.

In an RF range, a chip resistor must be represented by an ideal resistor R_p, in parallel with an ideal capacitor C_p, when the specified value of the chip resistor is higher than about 10 Ω. On the contrary, the chip resistor must be represented by an ideal resistor R_p, in parallel with an ideal inductor L_p when the specified value of the chip resistor is lower than about 10 Ω. In an RF range its model is not an *LC* resonant circuit, so there is no self-resonant frequency. This is a good approximation when the chip resistor is operating in an RF range, though actually the chip resistor has a resonant frequency in a microwave range which is much higher than a usual RF frequency.

4.2 Characterization of a Chip Part by Network Analyzer

An effective means of characterizing of a chip part is the S parameter measurement by a network analyzer as shown in Figure 4.2.

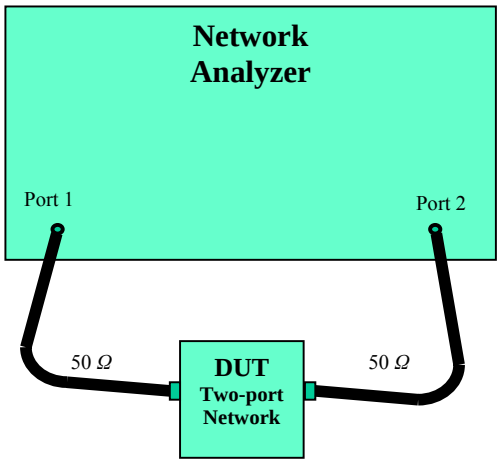


Figure 4.2 Characterization of a chip capacitor C, chip inductor L, chip resistor R by a network analyzer

For a chip capacitor or a chip inductor, its equivalent circuit consists of all three passive parts, a capacitor, an inductor, and a resistor. It is hard to distinguish the capacitance and inductance from a single-port testing, that is, from the S_{11} or S_{22} testing, because the reading of reactance is a mixed result of the capacitance and the inductance. Therefore, to characterize either chip capacitor or chip inductor it is more convenient to use two-port testing, that is, by S_{21} testing. However, for chip resistor, the equivalent circuit in an RF range contains only two parts, either a resistor and a capacitor or a resistor and an inductor. It is enough to characterize it by the single port testing, from either S_{11} or S_{22} testing. Figures 4.3, 4.4, and 4.5 show the set-ups for S-parameter testing of a chip capacitor, inductor, and resistor, respectively.

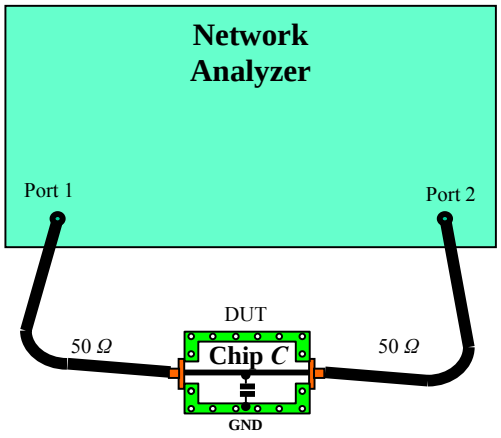


Figure 4.3 Two port S-parameter testing for a chip capacitor.

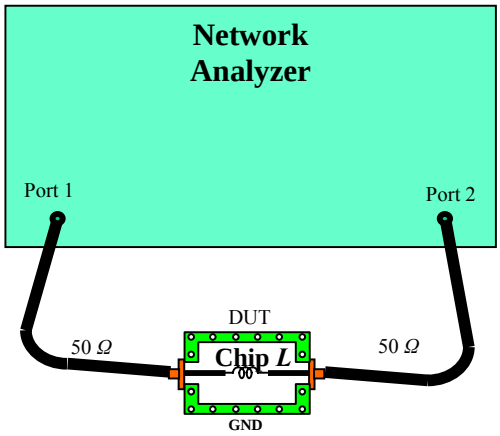


Figure 4.4 Two port S-parameter testing for a chip inductor.

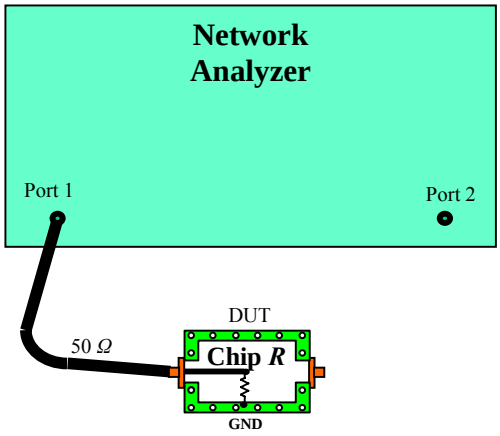


Figure 4.5 Single port S-parameter testing for a chip resistor.

4.3 Extraction from the Measurement by Network Analyzer

Two-port S-parameter measurements are for either S_{21} or S_{12}. The measurement for a chip C and chip L is symmetrical to port 1 and port 2. Therefore, we have

$$S_{21} = S_{12} \ . \tag{4.1}$$

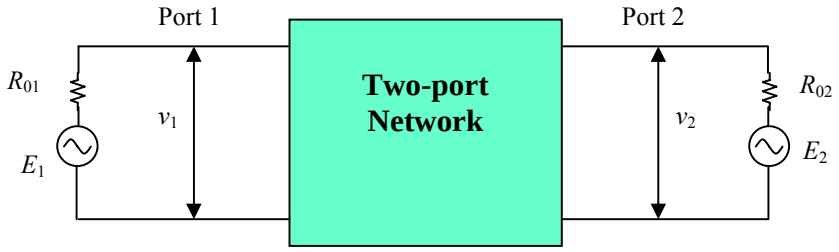


Figure 4.6 Source and load in two-port S-parameter measurement.

As shown in Figure 4.6, S_{21} or S_{21} can be expressed as a function of the input and output voltage, v_1 and v_2, input source and output load resistance, R_{o1} and R_{o2}, and input source and output load signal voltage, E_1 and E_2. This has been well derived in the book written by Ralph Carson (Ralph Carson, High Frequency Amplifier, 1975, pp. 152-155).

When $E_2 = 0$,

$$S_{21} = 2\sqrt{\frac{R_{01}}{R_{02}}}\frac{v_2}{E_1} \ , \tag{4.2}$$

Usually,

$$R_{01} = R_{02} = 50 \ \Omega, \tag{4.3}$$

then,

$$S_{21} = 2\frac{v_2}{E_1} \ . \tag{4.4}$$

The ratio v_2/E_1 is determined by the DUT (Device Under Test).

4.3.1 Extraction for Chip Capacitors

In the S_{21} measurement for chip capacitors as shown in Figure 4.3, Figure 4.6 can be replaced by Figure 4.7.

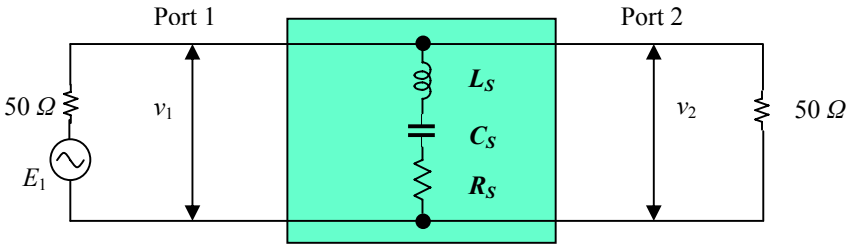


Figure 4.7 Source and load in two-port S-parameter measurement for a chip capacitor

By a simple but tedious mathematic derivation, it is easy to find

$$\frac{v_2}{E_1} = \frac{\sqrt{\left(R_S M + N^2\right)^2 + \left(MN - R_S N\right)^2}}{M^2 + N^2} \quad , \tag{4.5}$$

where

$$M = 50 + R_S \quad , \tag{4.6}$$

and

$$N = L_S \omega - \frac{1}{C_S \omega} \quad , \tag{4.7}$$

then,

$$S_{21} = 20\log\left(2\frac{v_2}{E_1}\right) = 20\log 2\frac{\sqrt{\left(R_S M + N^2\right) + \left(MN - R_S N\right)^2}}{M^2 + N^2} \quad . \tag{4.8}$$

At the self-resonant frequency,

$$\omega = \omega_{SRF} = \frac{1}{\sqrt{L_S C_S}} \quad , \tag{4.9}$$

$$N = 0 \quad , \tag{4.10}$$

$$S_{21,SRF} = 20\log\left(2\frac{v_2}{E_1}\right) = 20\log\left(2\frac{R_S}{M}\right) \quad , \tag{4.11}$$

or,

$$R_S = \frac{50}{2 \bullet 10^{-\frac{S_{21,SRF}}{20}} - 1} \quad , \tag{4.12}$$

$$L_S = \frac{1}{\omega_{SRF}^{\ 2} C_S} \quad , \tag{4.13}$$

$$C_S = C_{specified} \quad . \tag{4.14}$$

where $S_{21,SRF} = S_{21}$ at the self resonant frequency,
ω_{SRF} = Angular self resonant frequency,
$C_{specified}$ = Specified value of capacitance by the manufacturing.

From expressions (4.12), (4.13) and (4.14), it can be seen that the values of R_s and L_s can be obtained from the reading of $S_{21,SRF}$ at the self resonant frequency, ω_{SRF}.

Figure 4.8 shows a display on the screen of the network analyzer when S_{21} measurement for a chip capacitor is conducted.

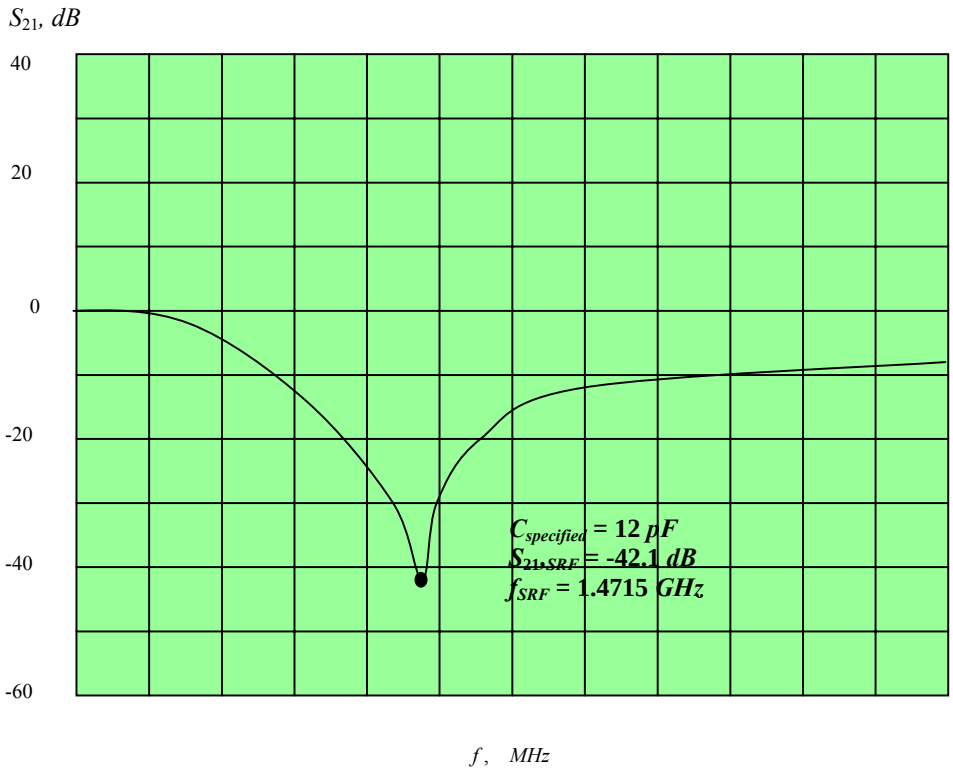


Figure 4.8 S_{21} plot for a chip capacitor testing

This is a test example for a chip capacitor with

$$C_{specified} = 12\ pF \ ,$$

and it is found that

$$\omega_{SRF} = 2\pi \cdot 1.4715\ GHz\ ,$$
$$S_{21,SRF} = -42.1\ dB.$$

In terms of expressions (4.12) and (4.13),

$$Ls = 0.97\ nH,$$
$$R_s = 0.20\ \Omega.$$

In RF circuit designs or for the RF circuit tests, it would be desirable to characterize all of chip capacitors. However, it would take up a tremendous amount of time to do so since there are so many different types of chip capacitors, all with different sizes and different manufacturers. Therefore, the test can only be conducted with a limited number of capacitor samples.

Figure 4.9 shows the self-resonant frequencies of various MuRata chip capacitors. The testing has been conducted for two different sizes: 30x60 $mils^2$ and 50x80 $mils^2$. The results of this sampling test where $C_{specified}$ =1.8 pF to 18,000 pF indicate that their equivalents for two different sizes are almost identical with each other. Figure 4.9 also illustrates the character of an LC resonant circuit in series, which appears as a capacitor by itself when it is operating under its self-resonant frequency. The bottom left area of Figure 4.9 below the f_{SRF} line is therefore called the "capacitive application region". On the contrary, it appears as an inductor when it is operating above its self-resonant frequency. The top right area of Figure 4.9 above the f_{SRF} line is thus called the "inductive application region".

In the logarithmic coordinates, all the tested points are crowded on a straight line. Consequently, the self-resonant frequency can be summarized by a simple equation, that is,

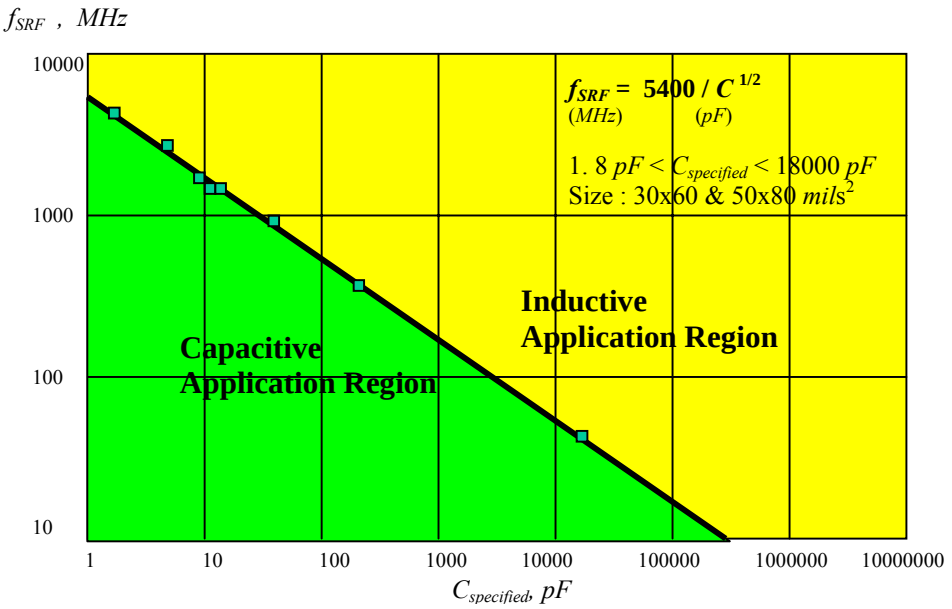


Figure 4.9 Plot of the self-resonant frequency f_{SRF} vs specified value of chip capacitor $C_{specified}$, extrated from testing of MuRata chip capacitors.

$$f_{SRF} = \frac{5400}{\sqrt{C_{specified}}} \ , \qquad (4.15)$$

where the unit of $C_{specified}$ is pF, and
the unit of f_{SRF} is MHz.

Equation (4.15) provides extreme convenience to the designers in the calculation of the self-resonant frequency of a chip capacitor. However, instead of equation (4.15), an experienced engineer should be familiar with the values of the special chip capacitors listed in Table 4.1.

Table 4.1 SRF (Self Resonant Frequency) of chip capacitors

SRF (MHz)	*Value of chip capacitor*, (pF)
40	**18,000**
50	**11,664**
100	**2,916**
150	**1,296**
450	**144**
500	**117**
800	**46**
900	**36**
1,000	**29**
1,500	**13**
2,400	**5.1**
5,400	**1.0**
6,235	**0.75**

Notes: It is impracticable to look for "zero" capacitors
1) when f < 40 *MHz*, because it is hard to find out the chip capacitors with C >18000 *pF* in the market, and
2) when f > 6235 *MHz*, because its relative tolerance of value is too high, and in addition, for low values of chip capacitors, there are only 0.75 and 0.5 *pF* available in the market.

Figure 4.10 shows the change of the in-series equivalent parasitic inductance *Ls* as $C_{specified}$ is varied, which is simply calculated from the self-resonant frequency via equations (4.13) and (4.14) for MuRata chip capacitors.

It is found that the in-series parasitic inductance *Ls* is kept almost constant from the low capacitance parts to high capacitance parts. Its average value is about 0.86 *nH*.

Finally from the reading of $S_{21,SRF}$, we can calculate the parasitic resistance in series, R_S , though the equation (4.12). Figure 4.11 plots its result. The tested result displays that it is scattered between 0.08 to 0.52 Ω when $C_{specified}$ is between 1.8 *pF* and 18,000 *pF*.

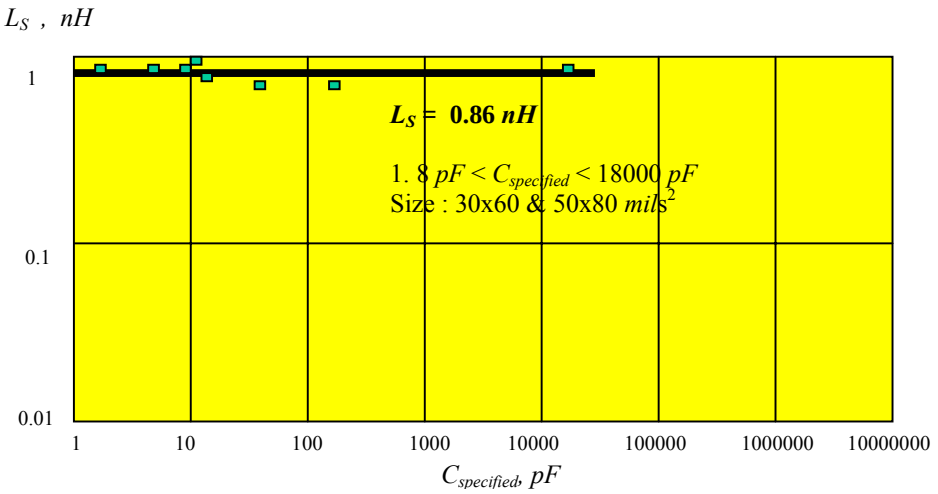


Figure 4.10 Plot of in-series parasitic inductance L_S vs specified capacitance $C_{specified}$, extracted from testing of MuRata chip capacitors.

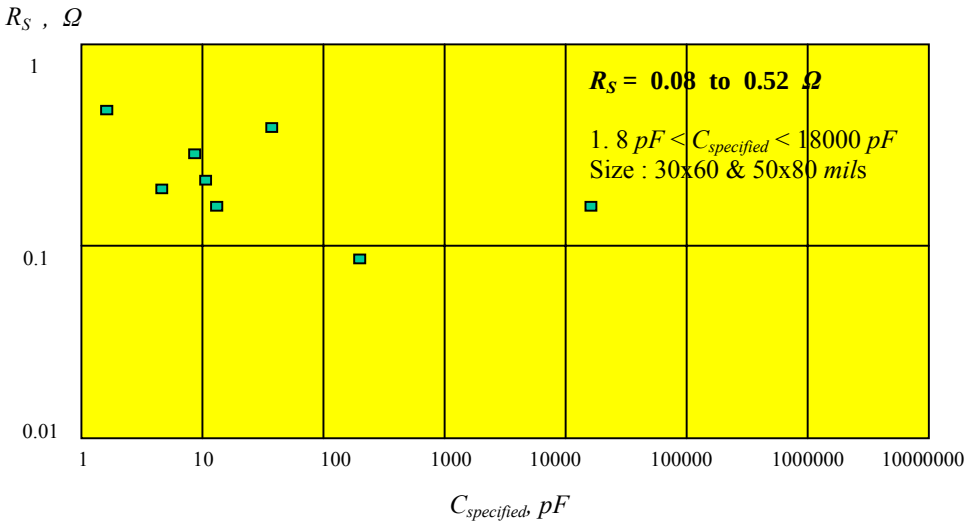


Figure 4.11 Plot of in-series resistance R_S vs specified capacitance $C_{specified}$, extrated from testing of MuRata chip capacitors.

4.3.2 Extraction for Chip Inductors

In the S_{21} measurement for chip inductors as shown in Figure 4.2, Figure 4.4 can be replaced by Figure 4.12.

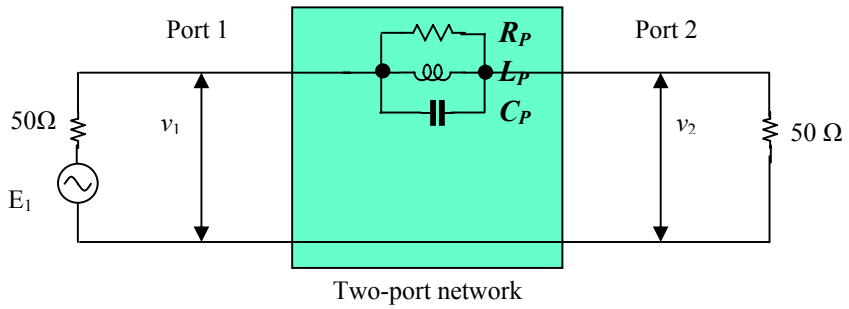


Figure 4.12 Source and load in a two-port S-parameter measurement for a chip inductor.

By a simple but tedious mathematic derivation, it is easy to find

$$\frac{v_2}{E_1} = \frac{50}{100 + R_p\left[1 + U\left(1 - \frac{1}{V}\right)^2\right]^{\frac{1}{2}}} \quad , \tag{4.16}$$

where

$$U = \left(R_P C_P \omega\right)^2 \quad , \tag{4.17}$$

$$V = L_P C_P \omega^2 \quad . \tag{4.18}$$

We have

$$S_{21} = 20\log\left(2\frac{v_2}{E_1}\right) = 20\log\frac{100}{100 + R_p\left[1 + U\left(1 - \frac{1}{V}\right)^2\right]^{\frac{1}{2}}} \quad . \tag{4.19}$$

At the self-resonant frequency,

$$\omega = \omega_{SRF} = \frac{1}{\left(L_P C_P\right)^{\frac{1}{2}}} \quad , \tag{4.20}$$

$$V = 1 \quad . \tag{4.21}$$

$$S_{21,SRF} = 20\log\frac{100}{100 + R_P} \quad , \tag{4.22}$$

or,

$$R_P = 100\left[10^{-\frac{S_{21,SRF}}{20}} - 1\right] \quad . \tag{4.23}$$

$$C_P = \frac{1}{\omega_{SRF}{}^2 L_P} \ . \tag{4.24}$$

$$L_P = L_{specified} \ . \tag{4.25}$$

where $S_{21,SRF} = S_{21}$ at the self-resonant frequency,
ω_{REF} = Angular self-resonant frequency,
$L_{specified}$ = Specified value of inductance by the manufacturing.

From expressions (4.23), (4.24) and (4.25), it can be seen that the values of R_P and C_P can be obtained from the reading of $S_{21,REF}$ at self-resonant frequency, ω_{SRF}.

Figure 4.13 shows a display on the screen of the network analyzer when S_{21} measurement for a chip inductor is conducted. This is a test example for a chip inductor with

$$L_{specified} = 60 \ nH,$$

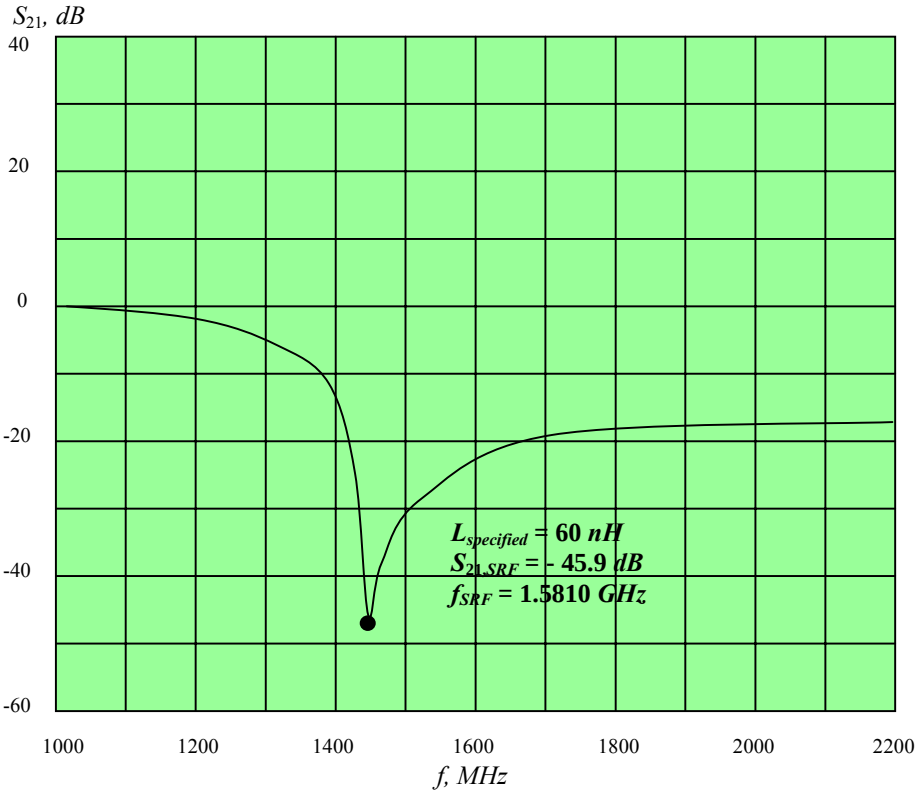


Figure 4.13 S_{21} plot for a chip inductor testing

and it is found that

$$\omega_{SRF} = 2\pi \cdot 1.5810 \ GHz,$$

$$S_{21,SRF} = -45.9\ dB.$$

In terms of expressions (4.23), (4.24) and (4.25),

$$C_P = 0.17\ pF,$$

$$R_p = 19624\ \Omega.$$

In RF circuit designs or RF circuit tests, it would be desirable to characterize all of the chip inductors. However, it would take up a tremendous amount of time to do so since there are so many different types of chip inductors, all with different sizes and varied manufacturers. Therefore, the tests can only be conducted with a limited number of inductors as samples.

Figure 4.14 shows the self-resonant frequencies of some MuRata chip inductors.

The testing has been conducted for one size: 80x120 $mils^2$. It is a sampling test from $L_{specified}$=22 nH to 1800 nH. Figure 4.14 illustrates the characteristics of the LC resonant circuit in parallel for a chip inductor, which appears as an inductor by itself when it is operating under its self-resonant frequency. The bottom left area of Figure 3.26 below the f_{SRF} line is therefore called the "inductive application region". On the contrary, it appears like a capacitor when it is operating above its self-resonant frequency. The top right area of Figure 4.14 above the f_{SRF} line is thus called the "capacitive application region".

In the logarithmic coordinates, all the tested points are crowded on a straight line. Consequently, the self-resonant frequency can be summarized by a simple equation, that is,

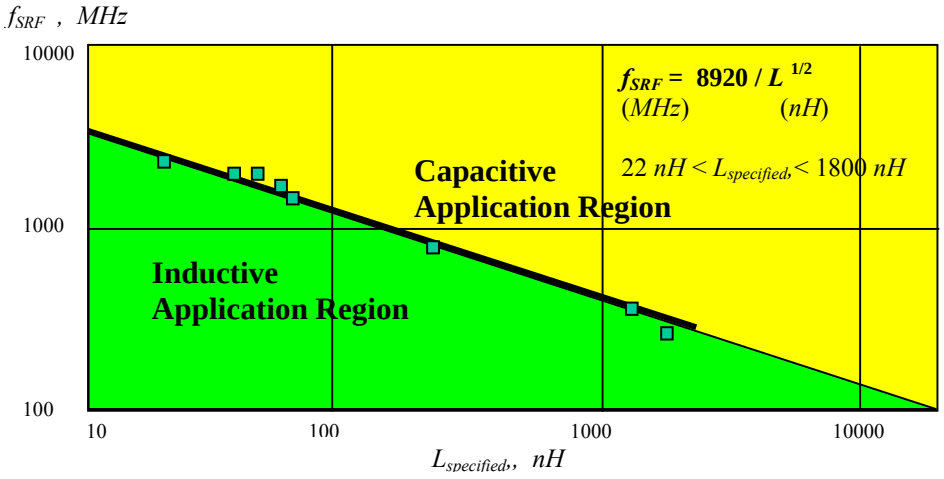


Figure 4.14 Plot of the self-resonant frequency f_{SRF} vs specified value of chip inductor $L_{specified}$, extracted from testing of MuRata chip inductors.

$$f_{SRF} = \frac{8920}{\sqrt{L_{specified}}}, \qquad (4.26)$$

where the unit of $L_{specified}$ is *nH*, and
the unit of f_{SRF} is *MHz*.

Equation (4.26) provides extreme convenience to the designers in the calculation of the self-resonant frequency of a chip inductor. Instead of equation (4.26), an experienced engineer should be familiar with the values of chip inductor listed in Table 4.2.

Table 4.2 SRF (Self Resonant Frequency) of chip inductors

SRF (MHz)	*Value of chip inductor*, (nH)
210	**1,800**
300	**884**
450	**393**
500	**318**
800	**124**
900	**98**
1,000	**79.6**
1,500	**35.4**
2,400	**13.8**
5,400	**2.7**
6,235	**2.0**

Notes: It is impracticable to look for "infinitive" inductors
1) when f < 210 *MHz*, because it is hard to find chip inductors with L >1800 *nH* in the market, and
2) when f > 6235 *MHz*, because its relative tolerance of value is too high and its accuracy is too low.

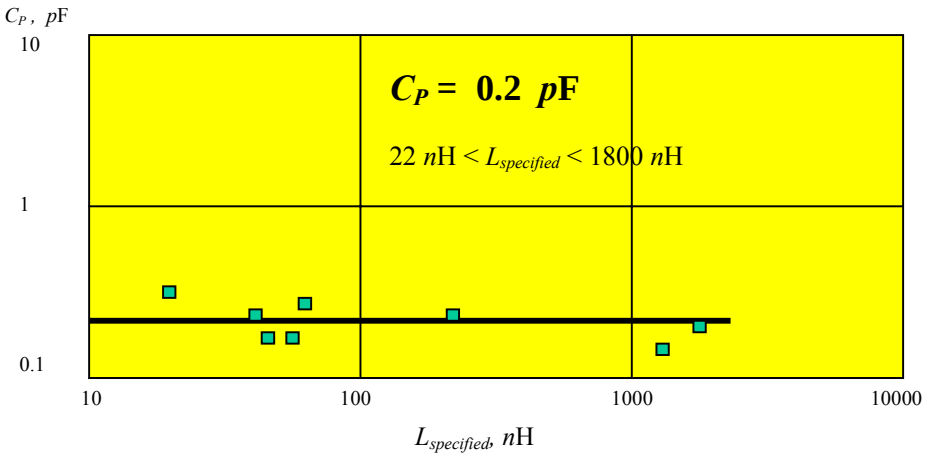


Figure 4.15 Plot of in-parallel capacitance C_P vs specified inductance $L_{specified}$, extracted from testing of MuRata chip inductors.

Figure 4.15 shows the change of the in-parallel equivalent parasitic capacitance C_P as $L_{specified}$ is varied, which is simply calculated from the self-resonant frequency via equation (4.24) for MuRata chip inductors,.

It is found that the in-parallel parasitic capacitance C_P is kept almost constant from low inductance parts to high inductance parts. Its average value is about 0.2 *pF*.

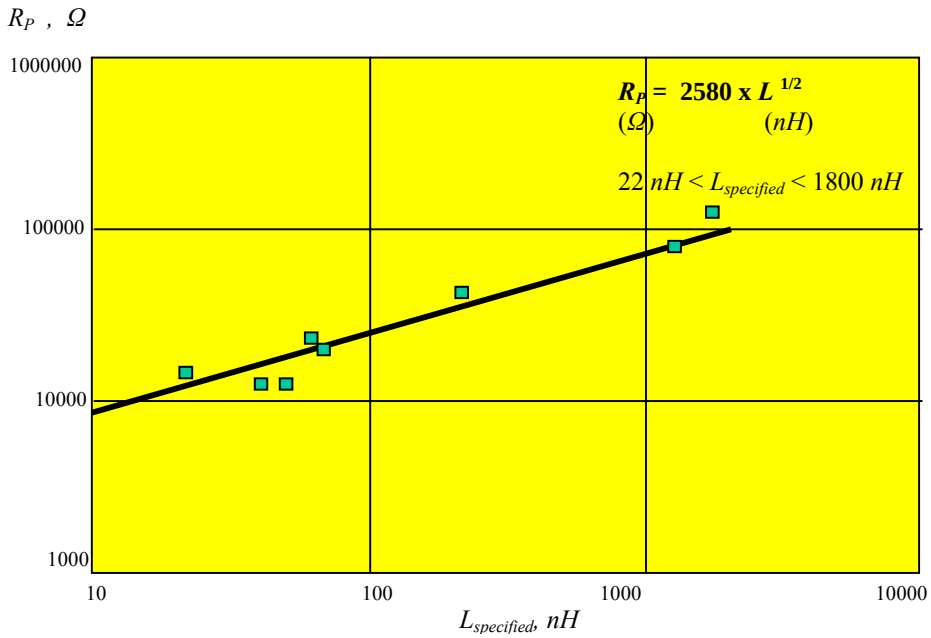


Figure 4.16 Plot of in-parallel resistance R_P vs specified inductance $L_{specified}$, extrated from testing of MuRata chip inductors.

Finally from the reading of $S_{21,SRF}$, we can calculate the in-parallel parasitic resistance R_P though equation (4.23). Figure 4.16 plots its result. The tested result shows that the in-parallel parasitic resistance R_P increases as the $L_{specified}$ is increased. In the logarithmic coordinates, all of tested point are crowded on a straight line. Consequently, the in-parallel parasitic resistance R_P can be summarized by a simple equation, that is,

$$R_P = \frac{2580}{\sqrt{L_{specified}}} \ , \qquad (4.27)$$

where the unit of $L_{specified}$ is *nH*, and
the unit of R_P is Ω.

Equation (4.27) provides extreme convenience to the designers in the calculation of the in-parallel parasitic resistance R_P of a chip inductor.

4.3.3 Extraction for Chip Resistors

In the S_{11} or S_{22} measurements of chip resistors, Figure 4.5 can be replaced by Figure 4.17.

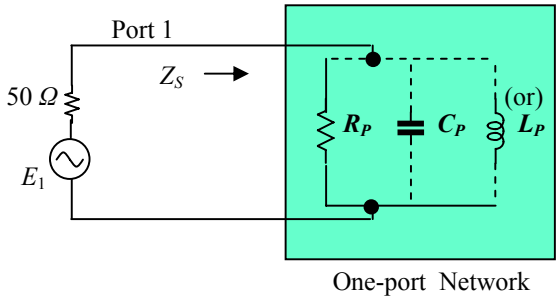


Figure 4.17 Single-port S-parameter measurement of a chip capacitor

The reading of S_{11} or S_{22} corresponds to a complex impedance Z_S, that is,

$$Z_S = R_S + jX_S = \frac{1+S_{11}}{1-S_{11}} \quad . \tag{4.28}$$

It should be noted that R_S and X_S are in series. In order to obtain R_P and C_P or L_P in parallel, the following conversion must be conducted:

$$Q = \frac{X_S}{R_S} \quad , \tag{4.29}$$

$$X_P = X_S\left(1+\frac{1}{Q_S{}^2}\right) \quad , \tag{4.30}$$

$$R_P = R_S\left(1+Q_S{}^2\right) \quad , \tag{4.31}$$

$$C_P = -\frac{1}{\omega X_P} \quad , \tag{4.32}$$

and,

$$L_P = -\frac{X_P}{\omega} \quad . \tag{4.33}$$

Figure 4.18 plots the chip resistor's resistance in parallel, R_P, versus the specified resistance $R_{specified}$.

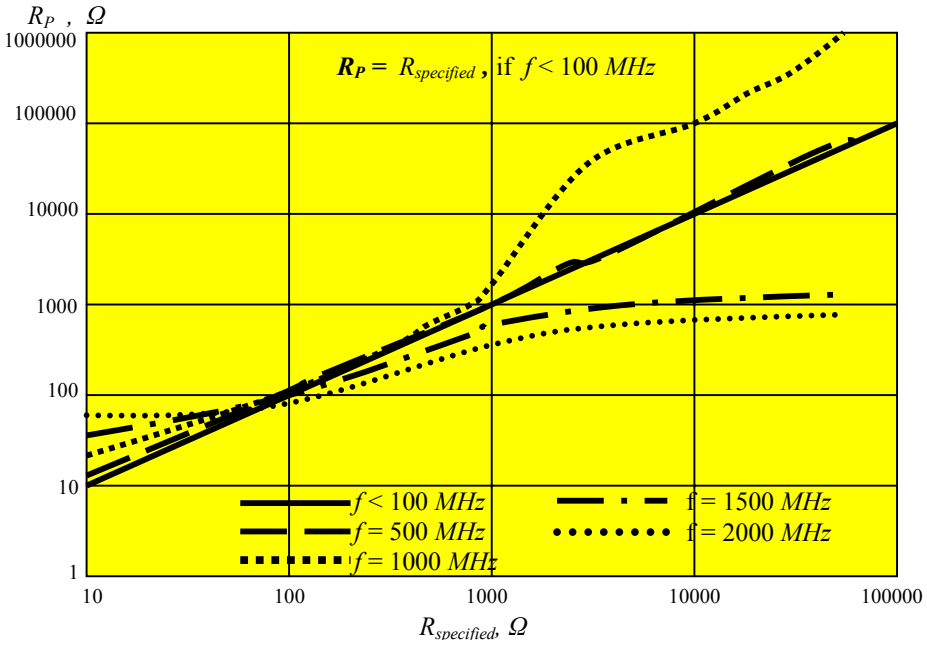


Figure 4.18 Plot of in parallel resistance R_P vs. specified resistance $R_{specified}$, extrated from testing of MuRata chip resistors.

Compared with a chip inductor or a chip capacitor, a special feature of a chip resistor is that R_P is a function of the operating frequency. As shown in Figure 4.18, R_P is almost always in the same value as $R_{specified}$ when the operating frequency is less than about 100 *MHz*. As the operating frequency increases, R_P deviates from the $R_{specified}$ more and more. At about 1000 *MHz*, the deviation approaches to a maximum. For example, a chip resistor with $R_{specified}$=10 k*Ω* becomes a resistor with $R_{specified}$=100 k*Ω*. The deviation is about 10 times. When the operating frequency is continuously increased, say, when the operating frequency is greater than about 1500 *MHz,* the deviation between R_P and $R_{specified}$ depends on the value of $R_{specified}$. *If* $R_{specified}$ >80*Ω,* the deviation is not continuously increased but decreased. The value of R_P is lower than that of $R_{specified}$. On the contrary, if $R_{specified}$ <80*Ω,* the deviation between R_P and $R_{specified}$ is increased so that the value of R_P is higher than that of $R_{specified}$. In the vicinity of $R_{specified}$ = 80 *Ω*, the deviation is negligible and R_P is almost independent from the operating frequency. A chip resistor with $R_{specified}$ = 80 *Ω* seems a good part to be applied in the practical circuit because it is neither sensitive to operating frequency nor deviated from its specified value.

Another special feature of a chip resistor is that its equivalent reactance is either C_P or L_P, depending on the value of $R_{specified}$. As shown in Figure 4.19, the reactance of the chip resistor is capacitive and the value of C_P is in the range of 0.65 to 0.90 *pF* if $R_{specified}$ is greater than 100 *Ω*. However, the reactance of the chip resistor is actually inductance if $R_{specified}$ is lower than about 6 *Ω*. This is quite an extraordinary feature.

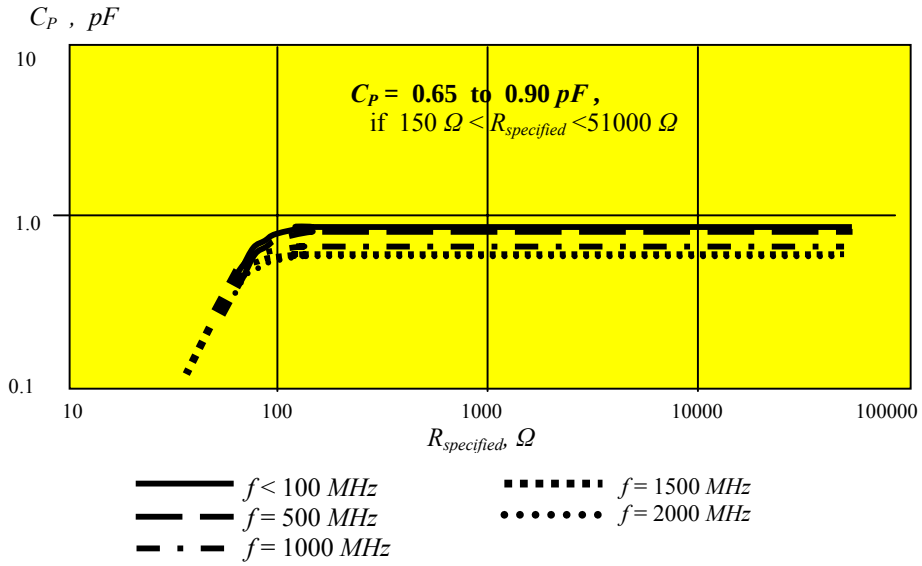


Figure 4.19 Plot of in-parallel capacitance C_P vs specified resistance $R_{specified}$, extrated from testing of MuRata chip resistors.

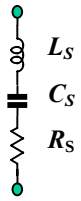
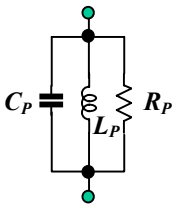
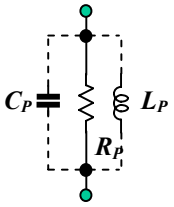
From Figure 4.19 it is also found that in the frequency range from 100 *MHz* to about 6000 *MHz*, the value of C_P is not too sensitive to the variation of the operating frequency though it is conceivable.

4. 4 Summary

Table 4.3 summarizes the equivalent circuits of the chip parts.

1) When a chip capacitor works at the frequency equal to its self-frequency, its impedance approaches to "zero". This chip capacitor is called a "zero" capacitor in respect to that frequency. By means of this character, the chip capacitor is widely applied as a by-pass capacitor in RF grounding.
2) When a chip inductor works at the frequency equal to its self-frequency, its impedance approaches infinity. By means of this character, the chip inductor is widely applied as a bias-fed choke in series to provide the DC power supply.
3) The self-resonant frequency of a chip resistor is very high. In an RF range, its resistance is the same as the specified value if the operating frequency is less than about 100 *MHz*. When the operating frequency is higher than about 100 *MHz*, its resistance can be deviated from the specified value more than 10 times and is a complicated function of the operating frequency. On the other hand, the chip resistor has an equivalent capacitor in parallel if its specified value is higher than about 6 *Ω*. However, instead of an equivalent capacitor, it has an equivalent inductor in parallel if its specified value is lower than about 6 *Ω*. Thus, the designer must be very careful to select chip resistors to be applied in RF circuits.

Table 4.3 Summary of the equivalent circuit of the chip parts.

Chip part	Equivalent circuit	Experical formula	Range
Capacitor ***C*** ($C=C_{specified}$)		$SRF=5400/C^{1/2}$ *MHz* $Cs=C$ *pF* $Ls=0.86$ *nH* $Rs=0.08$ to 0.52 *Ω*	$1.8\ pF < C < 18000\ pF$
Inductor ***L*** ($L=L_{specified}$)		$SRF=8920/L^{1/2}$ *MHz* $Lp=L$ *nH* $Cp=0.20$ *pF* $Rp=2580*L^{1/2}$ *Ω*	$22\ nH < L < 1800\ nH$
Resistor ***R*** ($R=R_{specified}$)		*Rp* (Refer to Figure 4.18) *Cp*=0.78 *pF* *Cp* or *Lp*? (Refer to Figure 4.19)	$f < 100\ MHz$ $100\ MHz < f < 2\ GHz$ $150\ Ω < R < 51\ kΩ$

References

[1] Ralph S. Carson, “High-Frequency Amplifiers,” (Book), , John Wiley & Sons, Inc., 1975.

[2] Manka W., “Alternative Methods for determining Chip Inductor Parameters,” IEEE Transactions on Parts, Hybrids, and Packaging, Vol. 13, Issue 4, pp. 378-385, December 1977.

[3] Lafferty, R., “Measuring the Self-Resonant Frequency of Capacitors,” IEEE Transactions on Components, Hybrids, and Manufacturing Technology, [see also IEEE Trans. on Components, Packaging, and Manufacturing Technology, Part A, B, C] , Vol. 5 , Issue 4 , pp. 528-530, December 1982.

[4] George D. Vendelin, Anthony M. Pavio, and Ulrich L. Rohde, “Microwave Circuit Design Using Linear and Nonlinear Techniques,” (Book), John Wiley & Sons, Inc., 1990.

[5] Paul R. Gray, Paul J. Hurst, Stephen H. Lewis, and Robert G. Meyer, “Analysis and Design of Analog Integrated Circuits,” (Book), Fourth Edition, John Wiley & Sons, Inc., 2001.

[6] Thomas H. Lee, “The Design of CMOS Radio-Frequency Integrated Circuits,” (Book), Cambridge University Press, 1998.

[7] Hannu Tenhunen, “CMOS Interconnects,” (Lecture), Electronic System Laboratory, Kungl Tekniska Hogskolan, 2000.

[8] Shim H.W., Zeeff, T.M., and Hubing T.H., “Decoupling strategies for printed circuit boards without power planes,” 2002 IEEE International Symposium on Electromagnetic Compatibility, , Vol. 1 , pp. 258-261, August 19-23, 2002.

Index

Contents

Chapter 5 Single-ended Stage and Differential Pair

5.1 Basic Single-ended Stage

5.1.1 General Description

The device is the core of a basic single-ended stage. It has 6 configurations:
1) CE (Common Emitter);
2) CB (Common Base);
3) CC (Common Collector, or Emitter Follower);
4) CS (Common Source);
5) CG (Common Gate);
6) CD (Common Drain, or Source Follower).

The former 3 configurations consist of bipolar devices while the latter 3 ones consist of MOSFETs. In digital circuits working at low and middle data rates, the designers very concern of their voltage and current gains. Table 5.1 lists the voltage and current gains for these 6 configurations.

Table 5.1 List of approximate voltage and current gain in a basic single-ended amplifier stage with low and mid-frequencies or low and mid-data rates

Configuration	Voltage gain, A_v	Current gain, A_i
CE	$-g_m(r_o // z_L)$	β_o
CB	$g_m z_L$	α_o
CC	$\dfrac{(\beta_o+1)\dfrac{r_o z_L}{r_o+z_L}}{r_b+r_\pi+(\beta_o+1)\dfrac{r_o z_L}{r_o+z_L}} \to 1$	$(1+g_m r_\pi)\dfrac{r_o}{r_o+z_L} \to (1+\beta_o)$
CS	$-g_m(r_d // z_L)$	$\to \infty$
CG	$(g_m+g_{mb})z_L$	$\to 1$
CD	$\dfrac{g_m r_d}{1+(g_m+g_{mb})r_d+\dfrac{r_d}{z_L}}$	$\to \infty$

From Table 5.1 it can be seen that all voltage gains depend on the load impedance z_L, while the current gains do not apparently depend on the load impedance z_L, except in the CC configuration. In the CE and CS configurations there is 180^o of phase difference between input and output voltage, and usually, their voltage gains are highest. This assertion conflicts with the expressions list in Table 5.1, in which the voltage gain in CE and CS configuration should be lower than that in the CB and CG configuration respectively. This is mainly due to the negligence of r_o, r_d, and other factors in the derivation of the voltage gains for the CB and CG configurations. In CC and CD configurations the voltage gain is about but less than 1 so that they are actually the lowest.

On the contrary, the current gain has a big difference among the various configurations. In CE and CC configurations, the current gains are about β_o. In CB and CG configurations, the current gains are about but less than 1, so that they are lowest, while in CS and CD configurations, the current gains approach to infinitive, the highest value. This is due to their high input impedances.

In the cases that a digital circuit is operating at a low data rate, it is enough to design the digital circuit if only voltage gain or current gain is considered. However, in RF or digital circuits at a high data rate, power transportation must be the first priority to be considered. Neither voltage nor current gain is more important than power gain here. The designers of these circuits very concern with the input and output impedances in various configurations, because the power gain is mainly determined by the performance of input and output impedance matching. The study of input and output impedances of a device with different configuration thus becomes a prerequisite in the designing of RF circuits or digital circuits at high data rates.

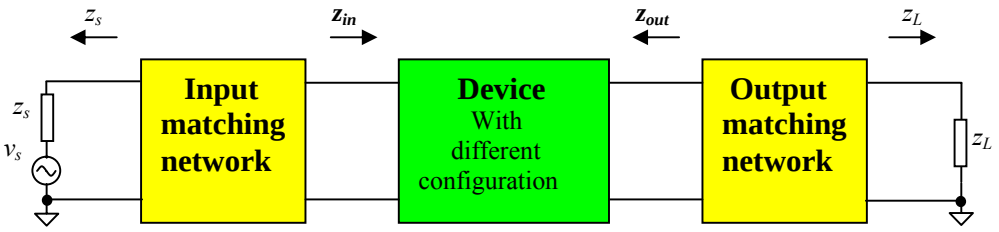
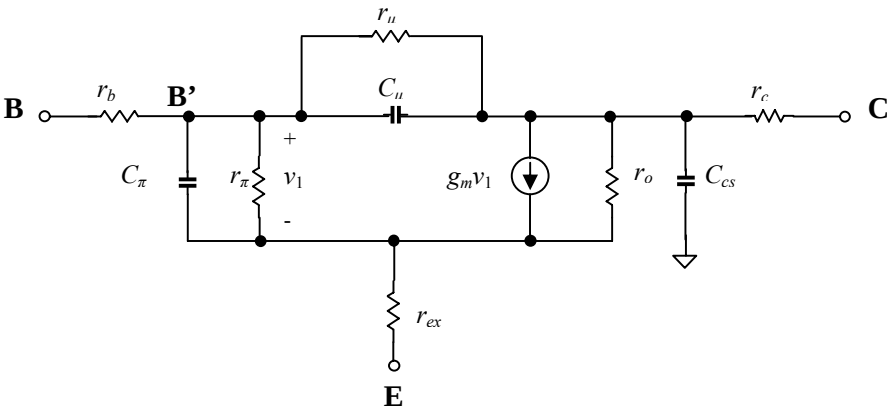


Figure 5.1 A basic single ended amplifier can be demarcated into 3 portions.

As a matter of fact, a basic single-ended amplifier stage can be demarcated into 3 portions as shown in Figure 5.1. The input matching network must be matched between the source impedance z_s and the input impedance of the device z_{in}, while the output matching network must be matched between the output impedance of the device z_{out} and the load impedance z_L. Consequently, the study of the input and output impedances z_{in} and z_{out} for the various configurations is the key issues in the entire design works for a single ended stage. In the following sub-sections, we will therefore focus on this topic on the basis of the equivalent circuit models of the devices.

5.1.2 Small Signal Model of a Bipolar Transistor

Figure 5.2 shows a small-signal equivalent circuit model of a bipolar transistor,



r_b = Contact resistance of base and the resistance between base and emitter,
C_π = Input capacitance, composed of the base-charging capacitance C_b and the emitter-base depletion layer capacitance C_{je} ,
r_π = Input resistance,
C_μ = Collector-base capacitance,
r_μ = Collector-base resistance,
r_{ex} = Emitter lead resistance,
C_{cs} = Collector-substrate capacitance,
r_o = Output resistance,
r_c = Collector resistance, composed of resistance from collector to buried layer, resistance with buried layer, and resistance from buried layer to base.

Figure 5.2 Small-signal equivalent circuit model of a bipolar transistor.

The DC characteristics of a bipolar transistor can be expressed by

$$I_c = I_s\left(1+\frac{V_{ce}}{V_A}\right)\exp\left(\frac{V_{be}}{V_T}\right) = I_s\left(1+\frac{V_{ce}}{V_A}\right)\exp\left(\frac{qV_{be}}{kT}\right) . \tag{5.1}$$

Its transconductance is

$$g_m = \frac{\partial I_c}{\partial V_{be}} = I_s\left(1+\frac{V_{ce}}{V_A}\right)\frac{q}{kT}\exp\left(\frac{qV_{be}}{kT}\right) = \frac{qI_c}{kT} \quad , \tag{5.2}$$

$$C_\pi = C_b + C_{je} \quad , \tag{5.3}$$

$$C_b = \tau_F g_m \quad , \tag{5.4}$$

$$C_{je} = \frac{C_{jeo}}{\sqrt[3]{1 - \frac{V_D}{\psi_o}}} \quad , \tag{5.5}$$

$$r_\pi = \frac{\beta_o}{g_m} \quad , \tag{5.6}$$

$$C_\mu = \frac{C_{\mu o}}{\left(1 - \frac{V_{cb}}{\psi_o}\right)^n} \quad , \tag{5.7}$$

$$r_\mu = (1 \to 10)\beta_o r_o \quad , \tag{5.8}$$

$$r_o = \frac{V_A}{I_c} \quad , \tag{5.9}$$

where g_m = Transconductance,
q = Charge of electron,
I_c = Total collector current,
I_s = Collector current when V_{be}=0,
k = Boltzmann constant,
T = Absolute temperature,
τ_F = Base transit time in the forward direction,
C_{jeo} = Emitter depletion-region capacitance when V_D = 0,
V_D = Bias on the junction,
Ψ_o = Built-in potential, a voltage across the junction when bias is zero,
β_o = Small-signal current gain,
$C_{\mu o}$ = The value of C_μ when V=0,
V_{ce} = Voltage drop from collector to base,
V_{be} = Voltage drop from base to emitter,
V_{cb} = Forward bias on the collector-base junction,
n = An exponent between 0.1 to 0.5,
V_A = Early voltage.

An understanding of the approximate range of each parameter's value in the model is very helpful in the circuit design. The emitter lead resistance r_{ex} is only a few Ω and is neglected if the emitter current is not too high. Usually, the output resistance r_o is about tens of kΩ while the collector resistance r_c is about 10 Ω to couple hundred Ω. The collector-base resistor r_μ has the highest value, and is usually in the range of tens of mega-Ω. The collector-base capacitance C_μ is very small and is usually only a few fF. The input capacitance C_π is about 0.1 to 2 pF while the collector-substrate capacitance C_{cs} is in the range of tens of fF.

Figure 5.3 and 5.4 are simplified from Figure 5.2 and correspond to the cases of high frequency and low/mid-frequency, respectively. The emitter lead resistance r_{ex} is neglected because it is usually quite small in respect to other resistances.

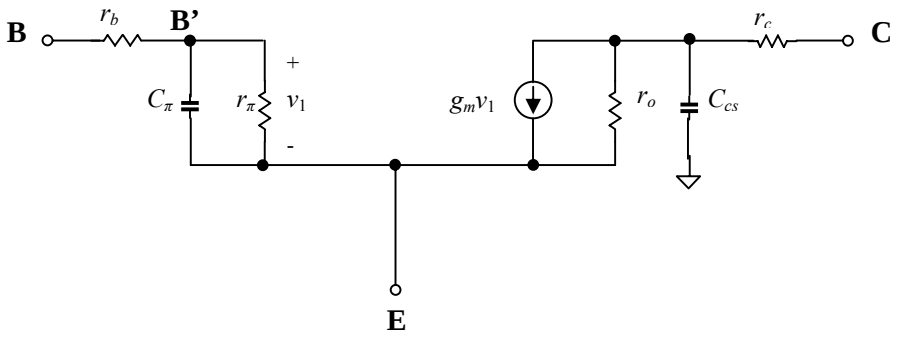


Figure 5.3 Small-signal equivalent circuit model of a bipolar transistor in a high frequency range.

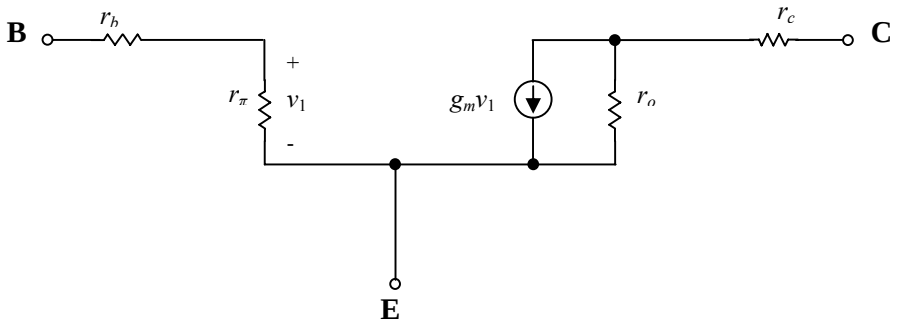


Figure 5.4 Small-signal equivalent circuit model of a bipolar transistor in low- and mid-frequency range.

5.1.2.1 Impedance of a CE (Common Emitter) Device

A common emitter amplifier is basically a bipolar transistor with its terminal E grounded as shown in Figure 5.5.

The input impedance is

$$z_{in} = r_b + \frac{1}{jC_\pi \omega} // r_\pi \quad . \tag{5.10}$$

Let's transfer the C_π and r_π in parallel to in series,

$$Q = \frac{1}{r_S C_S \omega} = r_\pi C_\pi \omega \ , \tag{5.11}$$

$$r_S = \frac{r_\pi}{\left(Q^2+1\right)} = \frac{r_\pi}{\left(r_\pi C_\pi \omega\right)^2 + 1} \ , \tag{5.12}$$

$$C_S = C_\pi \frac{Q^2+1}{Q^2} = C_\pi \frac{\left(r_\pi C_\pi \omega\right)^2 + 1}{\left(r_\pi C_\pi \omega\right)^2} \ . \tag{5.13}$$

Then,

$$z_{in} = \left[r_b + \frac{r_\pi}{\left(r_\pi C_\pi \omega\right)^2 + 1} \right] - j \frac{\left(r_\pi C_\pi \omega\right)^2}{C_\pi \omega [\left(r_\pi C_\pi \omega\right)^2 + 1]} \ . \tag{5.14}$$

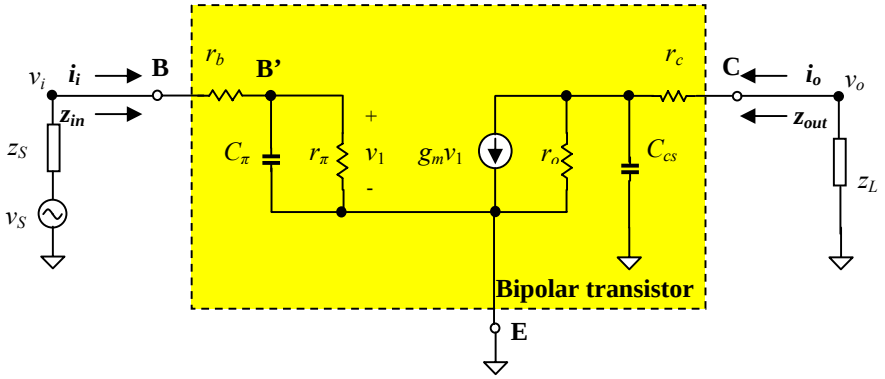


Figure 5.5 Small-signal equivalent circuit model of a CE (common emitter) device

On the other hand, the output impedance is

$$z_{out} = r_c + r_o \,//\, \frac{1}{jC_{cs}\omega} \ . \tag{5.15}$$

$$z_{out} = \left[r_c + \frac{r_o}{\left(r_o C_{cs} \omega\right)^2 + 1} \right] - j \frac{\left(r_o C_{cs} \omega\right)^2}{C_{cs} \omega [\left(r_o C_{cs} \omega\right)^2 + 1]} \ . \tag{5.16}$$

In the low and mid-frequency, the capacitors, C_π and C_{CS}, are negligible, therefore expressions (5.14) and (5.16) become

$$z_{in} = r_b + r_\pi \ , \tag{5.17}$$

and

$$z_{out} = r_c + r_o \ , \tag{5.18}$$

respectively.

5.1.2.2 Impedance of a CB (Common Base) Device

A common base amplifier is basically a bipolar transistor with its terminal B grounded as shown in Figure 5.6. In order to calculate the input impedance, Figure 5.6 is simplified to Figure 5.7 by the approximations of

$$r_b << r_\pi \quad , \tag{5.19}$$

and

$$r_\pi << r_o \quad , \tag{5.20}$$

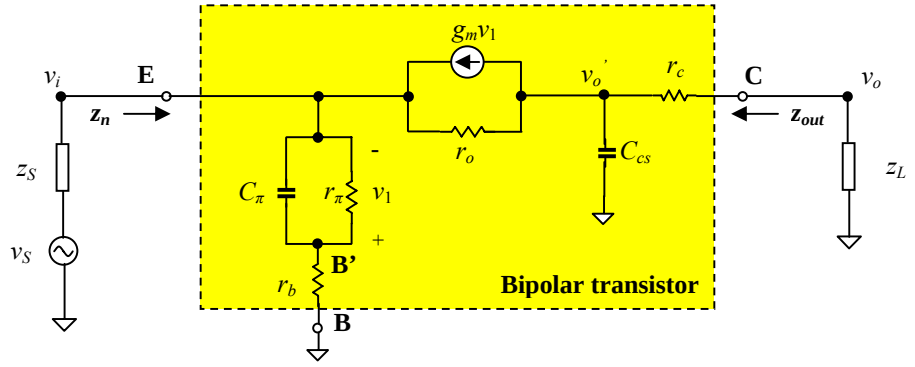


Figure 5.6 Small-signal equivalent circuit model of a CB (common base) device

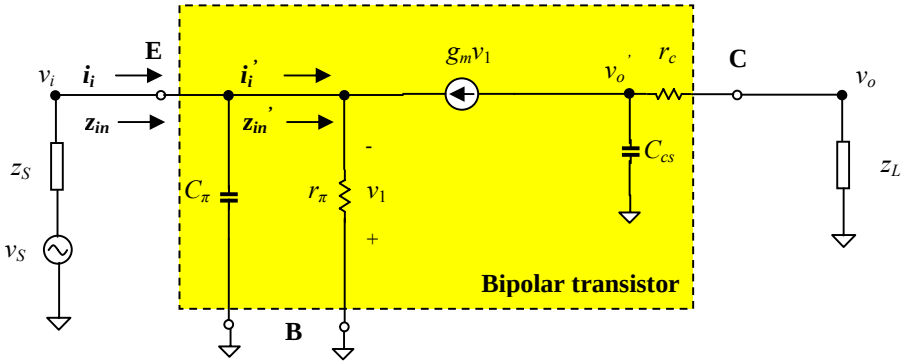


Figure 5.7 Small-signal equivalent circuit model of a CB (common base) device for input impedance discussion

Approximations (5.19) and (5.20) are usually reasonable in the actual devices. Under these approximations, r_b and r_o are removed from Figure 5.7, then,

$$v_i = -v_1 \quad , \tag{5.21}$$

$$z_{in}' = \frac{v_i}{i_i'} = \frac{v_i}{\frac{v_i}{r_\pi} - g_m v_1} = \frac{v_i}{\frac{v_i}{r_\pi} + g_m v_i} = \frac{r_\pi}{1 + g_m r_\pi} \approx \frac{1}{g_m} \quad , \tag{5.22}$$

and

$$z_{in} = \frac{1}{jC_\pi \omega} // z_{in}' = \frac{r_\pi}{1 + r_\pi (g_m + jC_\pi \omega)} \approx \frac{1}{g_m + jC_\pi \omega} \quad , \tag{5.23}$$

Equation (5.23) shows that the input impedance is independent of the load, z_L. However, for a higher accuracy when r_o is taken into account, the input impedance is dependent of the load, z_L. It would be modified to

$$z_{in} = \frac{r_\pi}{1 + r_\pi (g_m + jC_\pi \omega)} \left(1 + \frac{r_c + z_L}{r_o} \right) \quad , \tag{5.24}$$

In order to calculate the output impedance, Figure 5.6 is simplified to Figure 5.8, in which the r_b is neglected but r_o is kept.

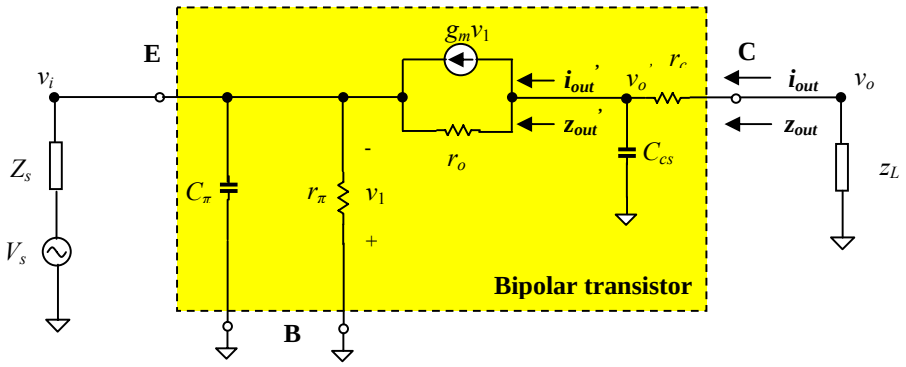


Figure 5.8 Small-signal equivalent circuit model of a CB (common base) device for output impedance discussion

At the node marked with v_o',

$$i_{out}' = g_m v_1 + \frac{v_o' - v_i}{r_o} = -g_m v_i + \frac{v_o' - v_i}{r_o} \quad , \tag{5.25}$$

and at the node E,

$$-\frac{v_i}{z_s // r_\pi // C_\pi}+\frac{v_o^{'}-v_i}{r_o}+g_m\left(-v_i\right)=0 \quad . \tag{5.26}$$

$$v_i=\frac{1}{r_o(\frac{1}{R}+\frac{1}{r_o}+g_m)}v_o^{'} \quad , \tag{5.27}$$

$$z_{out}^{'}=\frac{v_o^{'}}{i_{out}^{'}}=r_o+R(1+g_m r_o) \quad , \tag{5.28}$$

where

$$R=z_s // r_\pi // C_\pi=\frac{z_s r_\pi \frac{1}{jC_\pi\omega}}{z_s r_\pi+r_\pi\frac{1}{jC_\pi\omega}+z_s\frac{1}{jC_\pi\omega}},$$

$$R=\frac{z_s r_\pi}{z_s+r_\pi+jC_\pi\omega z_s r_\pi}, \tag{5. 29}$$

$$i_{out}=i_{out}^{'}+\frac{v_o^{'}}{\frac{1}{jC_{CS}\omega}}=\frac{v_o^{'}}{r_o+R(1+g_m r_o)}+v_o^{'}jC_{CS}\omega=\left[\frac{1}{r_o+R(1+g_m r_o)}+jC_{CS}\omega\right]v_o^{'} \quad , \tag{5.30}$$

$$v_o=i_{out}r_c+v_o^{'} \quad , \tag{5.31}$$

$$z_{out}=\frac{v_o}{i_{out}}=r_c+\frac{1}{\frac{1}{r_o+R(1+g_m r_o)}+jC_{CS}\omega} \quad , \tag{5.32}$$

At low- and mid-frequencies, the capacitors C_π and C_{CS} can be neglected so that equations (5.24) and (5.32) become

$$z_{in}=\frac{r_\pi}{1+r_\pi g_m}\left(1+\frac{r_c+z_L}{r_o}\right) \quad , \tag{5.33}$$

$$z_{out}=\frac{v_o}{i_{out}}=r_c+r_o+R(1+g_m r_o) \quad , \tag{5.34}$$

For a higher accuracy, the resistor r_b is taken into account, and after replacing of r_π in equation (5.33) by $(r_\pi + r_b)$, we have

$$z_{in} = \frac{r_\pi + r_b}{1 + (r_\pi + r_b) g_m} \left(1 + \frac{r_c + z_L}{r_o} \right) \quad , \tag{5.35}$$

where

$$R = \frac{z_s r_\pi}{z_s + r_\pi} . \tag{5.36}$$

It should be noted that the output impedance is dependent of the input impedance z_s which is implied in the intermediate parameter R and that the input impedance is also dependent of the load impedance z_L.

5.1.2.3 Impedance of a CC (Common Collector) Device

In a common collector amplifier (emitter follower), the terminal C is grounded. Figure 5.3 can be re-drawn as Figure 5.9. We then have

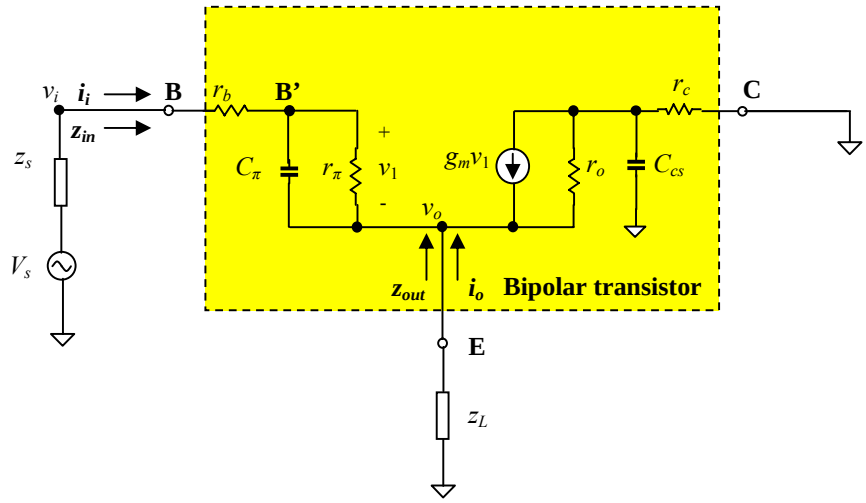


Figure 5.9 Small-signal equivalent circuit model of a CC (common collector) device

$$v_i = i_i (r_b + r_\pi // C_\pi) + v_o \quad , \tag{5.37}$$

$$v_1 = i_i (r_\pi // C_\pi) \quad , \tag{5.38}$$

$$v_o = (i_i + g_m v_1)[z_L // (r_o + r_c // C_{cs})] \quad . \tag{5.39}$$

Substitute (5.38) into (5.39),

$$v_o = i_i[1 + g_m(r_\pi // C_\pi)][z_L // (r_o + r_c // C_{cs})] \quad . \tag{5.40}$$

Replace v_o in (5.37) by (5.40),

$$v_i = i_i(r_b + r_\pi // C_\pi) + i_i[1 + g_m(r_\pi // C_\pi)][z_L // (r_o + r_c // C_{cs})] \quad . \tag{5.41}$$

Then,

$$z_{in} = \frac{v_i}{i_i} = (r_b + r_\pi // C_\pi) + [1 + g_m(r_\pi // C_\pi)][z_L // (r_o + r_c // C_{cs})] \quad . \tag{5.42}$$

Note that

$$r_\pi // C_\pi = \frac{1}{1 + jC_\pi \omega r_\pi} r_\pi \quad , \tag{5.43}$$

$$r_b + r_\pi // C_\pi = \frac{r_b + r_\pi + jC_\pi \omega r_b r_\pi}{1 + jC_\pi \omega r_\pi} \quad , \tag{5.44}$$

$$r_o + r_c // C_{cs} = \frac{r_o + r_c + jC_{cs} \omega r_o r_c}{1 + jC_{cs} \omega r_c} \quad , \tag{5.45}$$

$$z_L // (r_o + r_c // C_{cs}) = \frac{r_o + r_c + jC_{cs} \omega r_o r_c}{z_L + r_o + r_c + jC_{cs} \omega r_c (r_o + Z_L)} z_L \quad . \tag{5.46}$$

Then,

$$z_{in} = \frac{r_\pi + r_b(1 + jC_\pi \omega r_\pi) + (1 + \beta + jC_\pi \omega r_\pi) \dfrac{r_o + r_c + jC_{cs} \omega r_o r_c}{z_L + r_o + r_c + jC_{cs} \omega r_c (r_o + Z_L)} z_L}{1 + jC_\pi \omega r_\pi} \quad . \tag{5.47}$$

The output current consists of three portions when an output voltage is applied to the terminal E. The first portion is the current flowing through the r_o and r_c to the AC ground due to the output voltage. The second portion is the current generated by device, $g_m v_1$, in which v_1 is the voltage drop on r_π due to the output voltage. The third portion is the current flowing through r_π, r_b, and z_s to the ground due to the output voltage. We then have

$$i_o = \frac{v_o}{r_o + (r_c // C_{cs})} + g_m \frac{(r_\pi // C_\pi) v_o}{z_s + r_b + (r_\pi // C_\pi)} + \frac{v_o}{z_s + r_b + (r_\pi // C_\pi)} \quad , \tag{5.48}$$

$$i_o = \frac{1 + jC_{cs}\omega r_c}{r_c + r_o\left(1 + jC_{cs}\omega r_c\right)} v_o + \frac{1 + \beta + jC_\pi \omega r_\pi}{r_\pi + \left(Z_s + r_b\right)\left(1 + jC_\pi \omega r_\pi\right)} v_o \quad , \tag{5.49}$$

$$\frac{1}{z_{out}} = \frac{i_o}{v_o} = \frac{1 + jC_{cs}\omega r_c}{r_c + r_o\left(1 + jC_{cs}\omega r_c\right)} + \frac{1 + \beta + jC_\pi \omega r_\pi}{r_\pi + \left(z_s + r_b\right)\left(1 + jC_\pi \omega r_\pi\right)} \quad , \tag{5.50}$$

$$z_{out} = \frac{\left[r_c + r_o\left(1 + jC_{cs}\omega r_c\right)\right]\left[r_\pi + \left(z_s + r_b\right)\left(1 + jC_\pi \omega r_\pi\right)\right]}{\left(1 + jC_{cs}\omega r_c\right)\left[r_\pi + \left(z_s + r_b\right)\left(1 + jC_\pi \omega r_\pi\right)\right] + \left(1 + \beta + jC_\pi \omega r_\pi\right)\left[r_c + r_o\left(1 + jC_{cs}\omega r_c\right)\right]} \tag{5.51}$$

At a low- and mid-frequency range, the capacitors C_π and C_{CS} could be neglected, thus we have

$$z_{in} \approx r_\pi + r_b + \left(1 + \beta\right)\frac{r_o + r_c}{z_L + r_o + r_c} z_L \quad . \tag{5.52}$$

$$z_{out} \approx \frac{\left(r_c + r_o\right)\left(z_s + r_b + r_\pi\right)}{\left(1 + \beta\right)\left(r_c + r_o\right) + \left(z_s + r_b + r_\pi\right)} \quad . \tag{5.53}$$

It can be seen that the input impedance z_{in} is dependent of output resistors and load impedance z_L. Similarly, the output impedance z_{out} is dependent of input resistors and source impedance z_s.

A further approximation is that

if $$r_o >> z_L \quad , \tag{5.54}$$

then $$z_{in} \approx r_\pi + r_b + \left(1 + \beta\right) z_L \quad , \tag{5.55}$$

and if $$r_\pi >> z_s \quad , \tag{5.56}$$

then $$z_{out} \approx \frac{1}{1 + \beta + \dfrac{r_\pi}{r_c + r_o}} r_\pi \quad , \tag{5.57}$$

furthermore if

$$r_o >> r_\pi \quad , \tag{5.58}$$

then $$z_{out} \approx \frac{1}{g_m} \quad . \tag{5.59}$$

5.1.2.4 Comparison between CE, CB, and CC Device

It can be seen that the input and output impedances of a bipolar device are quite complicated and their complicated corresponding expressions must be adapted in the actual designs of RF or digital circuits at a high data rate.

In order to better understand the impedances' features let's take a look at these impedances in a low and mid-frequency range, which are much more simple than those in an RF frequency range or at a high data rate.

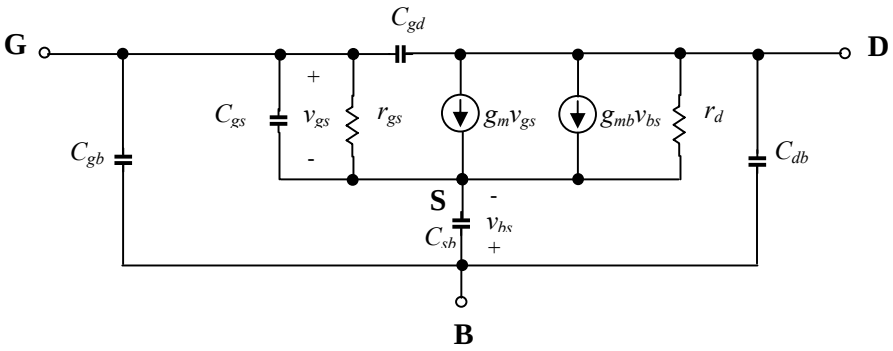
Table 5.2 lists the input and output impedances of a bipolar amplifier with a CE, CB, or CC configuration in a low and mid-frequency range. The resistors r_π and r_b play the main role to the input impedance while the resistors r_o and r_c play the main role to the output impedance. The CC amplifier has the highest input impedance and lowest output impedance. The CB amplifier has the lowest input impedance and highest output impedance. The input and output impedances of a CE amplifier are in intermediate values between CC and CB. Another remarkable feature is that the input and output impedances in a CB or CC configuration depend on the source impedance z_s and load impedance z_L while the input and output impedances in the CE configuration do not depend on the source impedance z_s and load impedance z_L. Especially, in the CB and CC configurations, the input impedance depends on the load impedance whereas the output impedance depends on the source impedance.

Table 5.2 Input and output impedances of a bipolar amplifier with a CE,CB, and CC configuration in a low- and mid-frequency range.

	Input impedance	*Output impedance*
CE (Common emitter)	$z_{in} = r_\pi + r_b$	$z_{out} = r_o + r_c$
CB (Common base)	$z_{in} \approx \frac{r_\pi + r_b}{1+(r_\pi + r_b)g_m}\left(1+\frac{r_c + z_L}{r_o}\right)$	$z_{out} \approx r_o + r_c + \frac{z_s r_\pi}{z_s + r_\pi}(1+g_m r_o)$
CC (Common collector, or Emitter follower)	$z_{in} \approx r_\pi + r_b + (1+\beta)\frac{r_o + r_c}{z_L + r_o + r_c} z_L$	$z_{out} \approx \frac{(r_o + r_c)(z_s + r_b + r_\pi)}{(1+\beta)(r_o + r_c)+(z_s + r_b + r_\pi)}$

5.1.3 Small-signal Model of a MOSFET

Figure 5.10 shows a small-signal equivalent circuit model of a MOSFET.



C_{gs} = Gate-source capacitance,
C_{gd} = Gate-drain capacitance,
C_{db} = Drain-body depletion capacitance,
C_{gb} = Gate-body capacitance,
C_{sb} = Source-body depletion capacitance,
r_d = Channel dynamic resistance,
r_{gs} = Gate-source resistance,
g_m = Top-gate transconductance,
g_{mb} = Body-effect transconductance.

. Figure 5.10 Small-signal equivalent circuit model of a MOSFET transistor

The DC characteristics of a MOSFET can be expressed by

$$I_d = \frac{k'}{2}\frac{W}{L}(V_{gs} - V_t)^2(1 + \lambda V_{ds}) \quad . \tag{5.60}$$

Its transcondutance is

$$g_m = \frac{\partial I_d}{\partial V_{gs}} = k'\frac{W}{L}\left(V_{gs} - V_t\right)\left(1 + \lambda V_{ds}\right) \quad , \tag{5.61}$$

if

$$\lambda V_{ds} << 1 \tag{5.62}$$

$$I_d = \frac{k'}{2}\frac{W}{L}(V_{gs} - V_t)^2 \quad , \tag{5.63}$$

$$g_m \approx k' \frac{W}{L}\left(V_{gs} - V_t\right) = \sqrt{2k' \frac{W}{L} I_d} \quad , \tag{5.64}$$

$$g_{mb} = \chi = \frac{C_{js}}{C_{ox}} \quad , \tag{5.65}$$

$$r_o = \frac{1}{\lambda I_d} = \frac{V_A}{I_d} \quad , \tag{5.66}$$

$$C_{gs} = \frac{2}{3} WLC_{ox} \quad , \tag{5.67}$$

$$f_t = \frac{1}{2\pi} \frac{g_m}{C_{gs} + C_{gd} + C_{gb}} \quad , \tag{5.68}$$

where I_d = Drain current forward active,
$k' = \mu_n C_{ox}$, and μ_n is the average electron mobility in the channel,
W = Channel width,
L = Channel length,
V_t = Threshold voltage,
V_{ds}= Voltage from drain to source,
V_{gs}= Voltage from gate to source,
V_A= Early voltage,
C_{js} = Capacitance per unit area of the depletion region under the channel,
C_{ox} = Oxide capacitance,
λ = Channel length modulation parameter,
χ = Ratio of g_m and g_{mb},
f_t = Transition frequency.

The gate-body capacitance C_{gb} is only a few fFs and can be neglected if this part is not too critical to the performance. The gate-source capacitance C_{gs} is about tens to hundreds of fFs and dominates over other capacitances. It must be taken good care of if the operating frequency is high. For simplicity the gate-drain capacitance C_{gd}, which is usually smaller than C_{gs}, is neglected. The drain-body and source-body capacitance C_{db} and C_{sb} are important because they have values in the same order as C_{gs} though they are usually smaller than C_{gs}. The value of top-gate transconductance g_m is usually a couple times higher than that of the body-effect transconductance g_{mb}. And, finally, the value of output resistance, r_o, is in the order of 10^5 Ω.

Figures 5.11 and 5.12 are simplified from Figure 5.10 and correspond to cases of high frequency and low/mid-frequency respectively. The source S is connected to the body B so that the current generator $g_{mb}V_{bs}$ is removed.

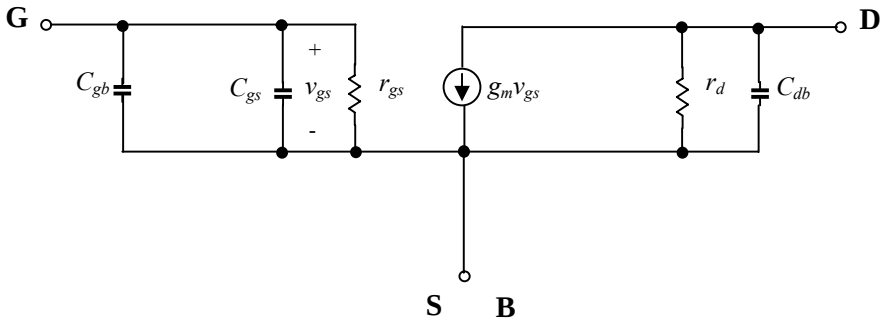


Figure 5.11 Small-signal equivalent circuit model of a MOSFET in a high frequency range (The source S is connected to the body B.)

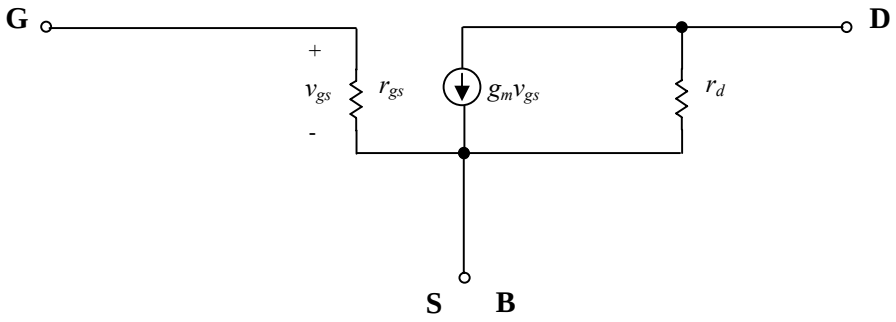


Figure 5.12 Small-signal equivalent circuit model of a MOSFET in a low mid-frequency range (The source S is connected to the body B.)

From the comparison of Figure 5.11 and 5.12 with Figures 5.3 and 5.4, it can be found that the topologies of equivalent circuit models between a bipolar transistor and a MOSFET are similar to each other. The 1st difference is that the bipolar transistor's input resistor r_b and output collector resistor r_c do not have corresponding parts in the MOSFET. The 2nd difference is that a collector-substrate capacitor C_{CS} appears in a bipolar transistor while a drain-body depletion capacitor C_{db} is found in a MOSFET.

The substrate in the bipolar collector-substrate capacitor C_{CS} is usually grounded whereas the substrate in the MOSFET drain-body depletion capacitor C_{db} can be grounded or connected by other ways. In our discussed MOSFET model we connected the substrate to the terminal S. Actually the substrate is grounded in the same way in the MOSFET CS (common source) case and the bipolar CE (common emitter) case. However, the topologies of C_{CS} and C_{db} are slightly different because in the MOSFET CG (common gate) and CD (common drain) cases the substrate is connected to terminal S while in the

bipolar CB (common base) and CC (common collector) cases the substrate is grounded. This slight difference might somehow impact the calculation of the input and output impedance but might be taken as a tolerance or an allowable approximation.

The similarity between bipolar and MOSFET models thus provides a short-cut to translate all of input and output impedance formula from the bipolar transistor to the MOSFET, as long as we do the following replacements:

$$
\begin{aligned}
r_b &\rightarrow 0\,, \\
r_c &\rightarrow 0\,, \\
r_\pi &\rightarrow r_{gs}\,, \\
C_\pi &\rightarrow (C_{gs}+C_{gb})\,, \\
r_o &\rightarrow r_d\,, \\
C_{cs} &\rightarrow C_{db}\,, \\
\beta &\rightarrow g_m r_{gs}\,.
\end{aligned}
$$

5.1.3.1 Impedance of a CS (Common Source) Device

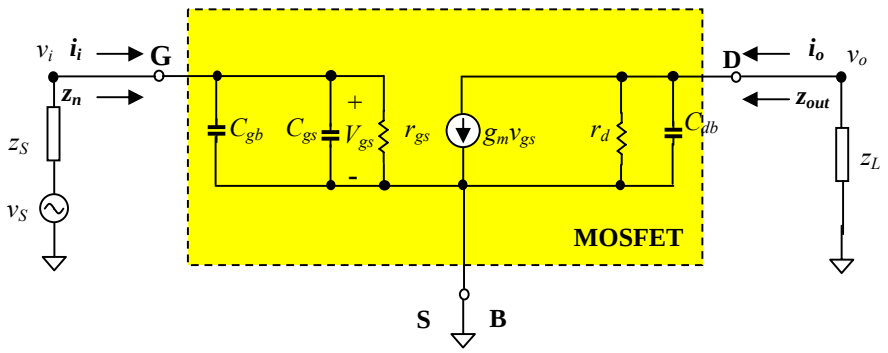


Figure 5.13 Small-signal equivalent circuit model of a CS (common source) device

Figure 5.13 shows the small-signal equivalent circuit model of a CS (common source) device. The input and output impedances are:

$$
z_{in} = \frac{r_{gs}}{[r_{gs}(C_{gs}+C_{gb})\omega]^2+1} - j\frac{[r_{gs}(C_{gs}+C_{gb})\omega]^2}{(C_{gs}+C_{gb})\omega\{[r_{gs}(C_{gs}+C_{gb})\omega]^2+1\}}\,. \tag{5.69}
$$

$$
z_{out} = \frac{r_d}{(r_d C_{db}\omega)^2+1} - j\frac{(r_d C_{db}\omega)^2}{C_{db}\omega[(r_d C_{db}\omega)^2+1]}\,. \tag{5.70}
$$

In a low- and mid-frequency range, all the capacitors could be neglected, thus we have

$$z_{in} = r_{gs} \quad , \tag{5.71}$$

$$z_{out} = r_d \quad . \tag{5.72}$$

5.1.3.2 Impedance of a CG (Common Gate) Device

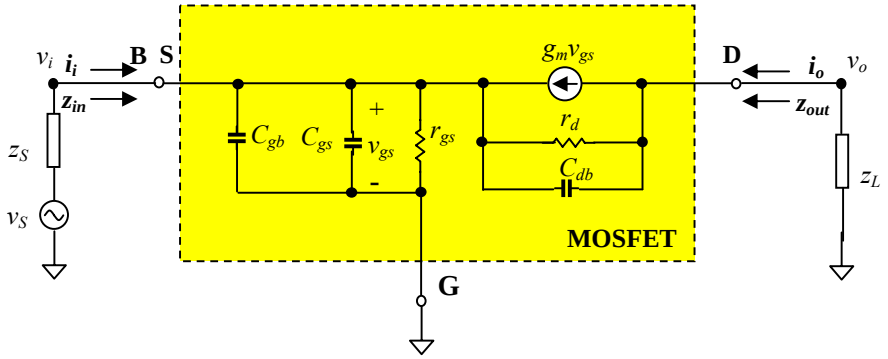


Figure 5.14 Small-signal equivalent circuit model of a CG (common gate) device

Figure 5.14 shows the small-signal equivalent circuit model of a CG (common gate) device. The input and output impedances are:

$$z_{in} = \frac{r_{gs}}{1 + r_{gs}[g_m + j(C_{gs} + C_{gb}\omega)]}\left(1 + \frac{z_L}{r_d}\right) \quad , \tag{5.73}$$

$$z_{out} = \frac{1}{\dfrac{1}{r_d + \dfrac{z_s r_{gs}}{z_s + r_{gs} + j(C_{gs} + C_{gb})\omega}(1 + g_m r_d)} + jC_{db}\omega} \quad . \tag{5.74}$$

In a low- and mid-frequency range, all the capacitors could be neglected, thus we have

$$z_{in} \approx \frac{r_{gs}}{1 + r_{gs} g_m}\left(1 + \frac{z_L}{r_d}\right) \quad , \tag{5.75}$$

$$z_{out} \approx r_d + \frac{z_s r_{gs}}{z_s + r_{gs}}(1 + g_m r_d) \quad . \tag{5.76}$$

5.1.3.3 Impedance of a CD (Common Drain) Device

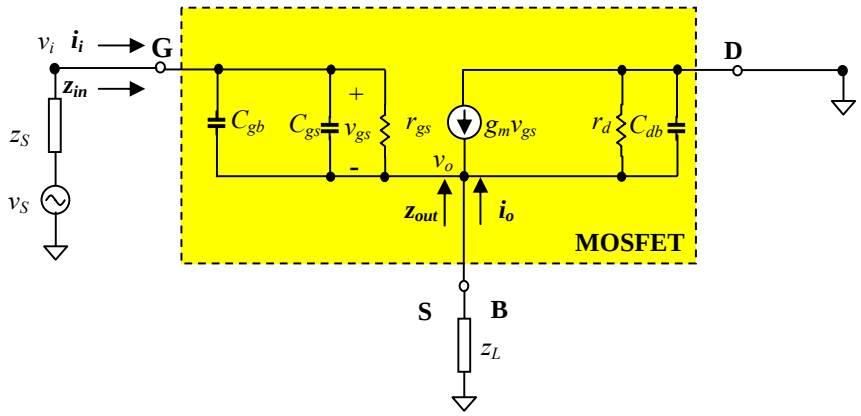


Figure 5.15 Small-signal equivalent circuit model of a CD (common drain) device

Figure 5.15 shows the small-signal equivalent circuit model of a CD (common drain) device. The input and output impedance are:

$$z_{in} = \frac{r_{gs} + [1 + g_m r_{gs} + j(C_{gs} + C_{gb})\omega r_{gs}]\dfrac{r_d}{z_L + r_d} z_L}{1 + j(C_{gs} + C_{gb})\omega r_{gs}} \quad , \tag{5.77}$$

$$z_{out} = \frac{r_d \{r_{gs} + z_s[1 + j(C_{gs} + C_{gb})\omega r_{gs}]\}}{r_{gs} + z_s[1 + j(C_{gs} + C_{gb})\omega r_{gs}] + r_d[1 + g_m r_{gs} + j(C_{gs} + C_{gb})\omega r_{gs}]} . \tag{5.78}$$

In a low and mid-frequency range, all the capacitors could be neglected, thus we have

$$z_{in} \approx r_{gs} + \left(1 + g_m r_{gs}\right)\frac{r_d}{z_L + r_d} z_L \quad , \tag{5.79}$$

$$z_{out} \approx \frac{r_d\left(z_s + r_{gs}\right)}{\left(1 + g_m r_{gs}\right) r_d + \left(z_s + r_{gs}\right)} \quad . \tag{5.80}$$

5.1.3.4 Comparison between CS, CG, and CD Device

It can be seen that the input and output impedances of a MOSFET are also quite complicated, and their complicated corresponding expressions must be adapted in an actual RF circuit design or in a high data rate circuit design.

In order to better understand the impedance features, let's take a look at these impedances in a low and mid-frequency range, which are much more simple than those in an RF range or at a high data rate.

Table 5.3 lists the input and output impedances of a MOSFET amplifier with a CS, CG, or CD configuration in a low/mid-frequency range. The resistors r_{gs} plays the main role to the input impedance while the resistors r_d plays the main role to the output impedance. The CD amplifier has the highest input impedance and lowest output impedance. The CG amplifier has lowest input impedance and highest output impedance. The input and output impedances of the CS amplifier are in intermediate values between CC and CB. Another remarkable feature is that the input and output impedances in CG and CD configurations depend on the source impedance z_s and load impedance z_L while the input and output impedances in the CS configuration do not depend on the source impedance z_s and load impedance z_L. Especially, in the CG and CD configurations, the input impedance depends on the load impedance whereas the output impedance depends on the source impedance.

Table 5.3 Input and output impedances of a MOSFET with a CS,CG, and CD configuration in a low- and mid-frequency range.

	Input impedance	*Output impedance*
CS (Common source)	$z_{in} = r_{gs}$	$z_{out} = r_d$
CG (Common gate)	$z_{in} \approx \frac{r_{gs}}{1 + r_{gs} g_m} \left(1 + \frac{z_L}{r_d} \right)$	$z_{out} \approx r_d + \frac{z_s r_{gs}}{z_s + r_{gs}} (1 + g_m r_d)$
CD (Common drain, or source follower)	$z_{in} \approx r_{gs} + \left(1 + g_m r_{gs}\right) \frac{r_d}{z_L + r_d} z_L$	$z_{out} \approx \frac{r_d \left(z_s + r_{gs} \right)}{\left(1 + g_m r_{gs}\right) r_d + \left(z_s + r_{gs} \right)}$

5.2 Differential Pair

5.2.1 DC Transfer Characteristic

5.2.1.1 DC Transfer Characteristic of a Bipolar Differential Pair

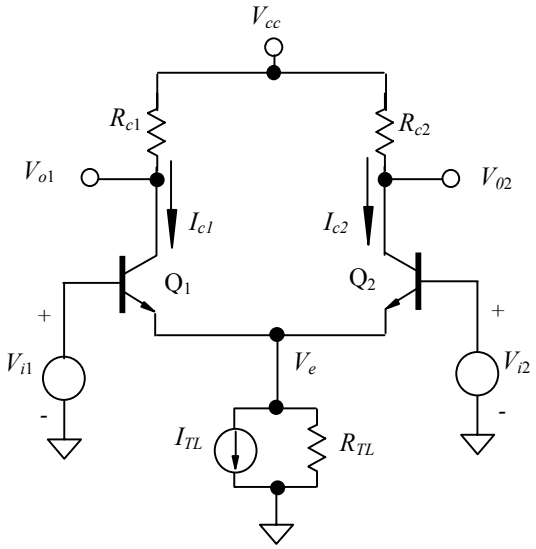


Figure 5.16 A differential pair with bipolar devices

Figure 5.16 shows a differential pair with bipolar devices. Its DC characteristics can be derived form the exponential relationship between the base voltage and the collector current. In the loop of V_{i1}→ emitter of Q_1→ V_e→ emitter of Q_2→ V_{i2},

$$V_{i1} - V_{be1} + V_{be2} - V_{i2} = 0\,, \tag{5.81}$$

From equation (5.1), we have

$$V_{be1} = V_T \ln \frac{I_{c1}}{I_{S1}\left(1 + \dfrac{V_{ce1}}{V_{A1}}\right)}\quad, \tag{5.82}$$

$$V_{be2} = V_T \ln \frac{I_{c2}}{I_{S2}\left(1 + \dfrac{V_{ce2}}{V_{A2}}\right)}\quad. \tag{5.83}$$

Assuming that the transistors are identical, that is

$$I_{S1} = I_{S2} = I_S\quad, \tag{5.84}$$

$$V_{ce1} = V_{ce2} = V_{ce} \quad , \tag{5.85}$$

$$V_{A1} = V_{A2} = V_A \quad , \tag{5.86}$$

$$\alpha_{F1} = \alpha_{F2} = \alpha_F \quad , \tag{5.87}$$

from expression (5.82) to (5.87), we have

$$\frac{I_{c1}}{I_{c2}} = \exp\left(\frac{V_{be1} - V_{be2}}{V_T}\right) = \exp\left(\frac{V_{i1} - V_{i2}}{V_T}\right) = \exp\left(\frac{V_{id}}{V_T}\right) \quad , \tag{5.88}$$

where $$V_{id} = V_{i1} - V_{i2} \quad . \tag{5.89}$$

By introducing of the tail current I_{TL}, we have

$$-\left(I_{e1} + I_{e2}\right) = I_{TL} = \frac{I_{c1} + I_{c2}}{\alpha_F} \quad , \tag{5.90}$$

from expressions (5.88) and (5.90), we have

$$I_{c1} = \frac{\alpha_F I_{TL}}{1 + \exp\left(-\frac{V_{id}}{V_T}\right)} \quad , \tag{5.91}$$

$$I_{c2} = \frac{\alpha_F I_{TL}}{1 + \exp\left(\frac{V_{id}}{V_T}\right)} \quad . \tag{5.92}$$

On the other hand,

$$V_{o1} = V_{cc} - I_{c1} R_c \quad , \tag{5.93}$$

$$V_{o1} = V_{cc} - I_{c1} R_c \quad , \tag{5.94}$$

if

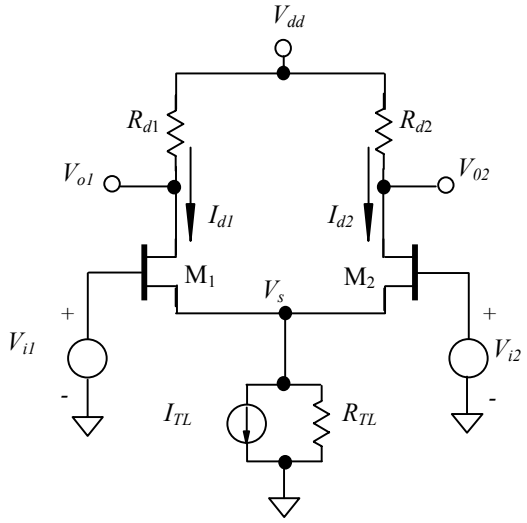
$$R_{c1} = R_{c2} = R_c \quad . \tag{5.95}$$

Finally,

$$V_{od} = V_{o1} - V_{o2} = \alpha_F I_{TL} R_c \tanh\left(\frac{-V_{id}}{2V_T}\right) . \tag{5.96}$$

Instead of a linear relationship between differential output and differential input voltage, the ideal differential output voltage is a tangent-hyperbolic function of the differential input voltage. It approaches to a linear relationship when V_{id} is small.

5.2.1.2 DC Transfer Characteristic of a CMOS Differential Pair



Figue 5.17 A differential pair with MOSFET devices

Figure 5.17 shows a differential pair with MOSFET devices. Its DC characteristics can be derived form the square-law relationship between the gate voltage and the drain current. In the loop of V_{i1}→ source of M_1→ V_s→ source of M_2→ V_{i2},

$$V_{i1} - V_{gs1} + V_{gs2} - V_{i2} = 0. \tag{5.97}$$

From equation (5.60), we have

$$V_{gs1} - V_t = \sqrt{\frac{2I_{d1}}{k'\frac{W_1}{L_1}(1+\lambda V_{ds1})}}\quad , \tag{5.98}$$

$$V_{gs2} - V_t = \sqrt{\frac{2I_{d2}}{k'\frac{W_2}{L_2}(1+\lambda V_{ds2})}}\quad . \tag{5.99}$$

Assuming that the transistors are identical, that is

$$W_1 = W_2 = W\quad , \tag{5.100}$$

$$L_1 = L_2 = L\quad , \tag{5.101}$$

$$V_{ds1} = V_{ds2} = V_{ds}\quad , \tag{5.102}$$

then, from (5.97) to (5.102), we have

$$V_{id} = V_{i1} - V_{i2} = V_{gs1} - V_{gs2} = \frac{\sqrt{I_{d1}} - \sqrt{I_{d2}}}{\sqrt{\frac{k'}{2}\frac{W}{L}(1+\lambda V_{ds})}} \quad . \tag{5.103}$$

On the other hand,

$$I_{d1} + I_{d2} = I_{TL} \quad , \tag{5.104}$$

from (5.103) and (5.104),

$$I_{d1} = \frac{I_{TL}}{2} + \frac{k'}{4}\frac{W}{L}(1+\lambda V_{ds})V_{id}\sqrt{\frac{4I_{TL}}{\frac{k'}{2}\frac{W}{L}(1+\lambda V_{ds})} - V_{id}^{\,2}} \quad , \tag{5.105}$$

$$I_{d2} = \frac{I_{TL}}{2} - \frac{k'}{4}\frac{W}{L}(1+\lambda V_{ds})V_{id}\sqrt{\frac{4I_{TL}}{\frac{k'}{2}\frac{W}{L}(1+\lambda V_{ds})} - V_{id}^{\,2}} \quad , \tag{5.106}$$

consequently,

$$\Delta I_d = I_{d1} - I_{d2} = \frac{k'}{2}\frac{W}{L}(1+\lambda V_{ds})V_{id}\sqrt{\frac{4I_{TL}}{\frac{k'}{2}\frac{W}{L}(1+\lambda V_{ds})} - V_{id}^{\,2}} \quad , \tag{5.107}$$

$$V_{od} = V_{o1} - V_{o2} = (V_{dd} - I_{d1}R_d) - (V_{dd} - I_{d2}R_d) = -\Delta I_d R_d \quad , \tag{5.108}$$

$$V_{od} = -R_d\frac{k'}{2}\frac{W}{L}(1+\lambda V_{ds})V_{id}\sqrt{\frac{4I_{TL}}{\frac{k'}{2}\frac{W}{L}(1+\lambda V_{ds})} - V_{id}^{\,2}} \quad . \tag{5.109}$$

Instead of a linear relationship between the differential output and input voltages, the differential output voltage is deviated from a linear relationship with the input differential voltage by the square root factor in expression (5.109). It approaches to a linear relationship when V_{id} is small.

5.2.2 Small Signal Characteristic

Figure 5.18 shows a differential pair with bipolar and MOSFET devices. It should be noted that this is not a simple combination but different from Figure 5.16 and 5.17. All the voltages and currents in Figure 5.16 and 5.17 are DC values which are denoted by

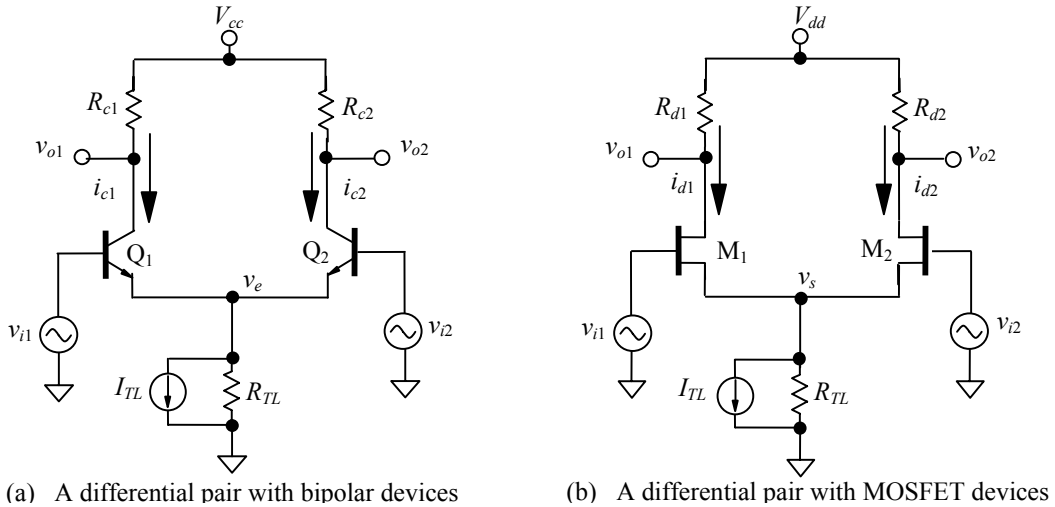
capital case letters while all the voltages and currents in Figure 4.18 are AC values which are denoted by lower case letters. Also, the two input sources in Figure 5.16 and 5.17 are DC sources while the two input sources in Figure 5.18 are AC sources.

In general, the voltage transfer characteristic of a differential pair can be expressed by

$$v_{o1} = A_{11}v_{i1} + A_{12}v_{i2} \quad , \tag{5.110}$$

$$v_{o2} = A_{21}v_{i1} + A_{22}v_{i2} \quad , \tag{5.111}$$

where
A_{11}= At port 1 output, voltage gain of v_{i1} at port 1 input, that is, v_{o1}/v_{i1}, when v_{i2}=0,
A_{12}= At port 1 output, voltage gain of v_{i2} at port 2 input, that is, v_{o1}/v_{i2} when v_{i1}=0,
A_{21}= At port 2 output, voltage gain of v_{i1} at port 1 input, that is, v_{o2}/v_{i1} when v_{i2}=0,
A_{22}= At port 2 output, voltage gain of v_{i2} at port 2 input, that is, v_{o2}/v_{i2}, when v_{i1}=0.



(a) A differential pair with bipolar devices (b) A differential pair with MOSFET devices

Figure 5.18 Differential pairs

In order to describe the special character of a differential pair, it is desirable to convert equations (5.110) and (5.111) into forms with common mode and differential mode. At the input of a differential pair, let's define an input differential mode voltage v_{id} and an input common mode voltage v_{ic},

$$v_{id} = v_{i1} - v_{i2} \quad , \tag{5.112}$$

$$v_{ic} = \frac{v_{i1} + v_{i2}}{2} \quad , \tag{5.113}$$

where v_{i1}= Input voltage at input port 1,
v_{i2}= Input voltage at input port 2,
v_{id}= Input differential mode voltage,
v_{ic}= Input common mode voltage.

Similarly, let's define an output differential mode voltage v_{od} and an output common mode voltage v_{oc} for the output of a differential pair,

$$v_{od} = v_{o1} - v_{o2} \quad , \tag{5.114}$$

$$v_{oc} = \frac{v_{o1} + v_{o2}}{2} \quad , \tag{5.115}$$

where v_{o1}= Output voltage at input port 1,
v_{o2}= Output voltage at input port 2,
v_{od}= Output differential mode voltage,
v_{oc}= Output common mode voltage.

The equations (5.110) and (5.111) can be converted to common mode and differential mode expressions in terms of expressions from (5.112) to (5.115), they are

$$v_{od} = A_{dm} v_{id} + A_{cm-dm} v_{ic} \quad , \tag{5.116}$$

$$v_{oc} = A_{dm-cm} v_{id} + A_{cm} v_{ic} \quad , \tag{5.117}$$

where
A_{dm}= Differential mode voltage gain, that is, v_{od}/v_{id} when v_{ic}=0,
$A_{cm\text{-}dm}$= Common mode to differential mode voltage gain, that is, v_{od}/v_{ic} when v_{id}=0,
$A_{dm\text{-}cm}$= Differential mode to common mode voltage gain, that is, v_{oc}/v_{id} when v_{ic}=0,
A_{cm}= Common mode voltage gain, that is, v_{oc}/v_{ic} when v_{id}=0.

The relationships of the various voltage gains in equations (5.110), (5.111) and in equations (5.116), (5.117) are

$$A_{dm} = \frac{A_{11} - A_{12} - A_{21} + A_{22}}{2} \quad , \tag{5.118}$$

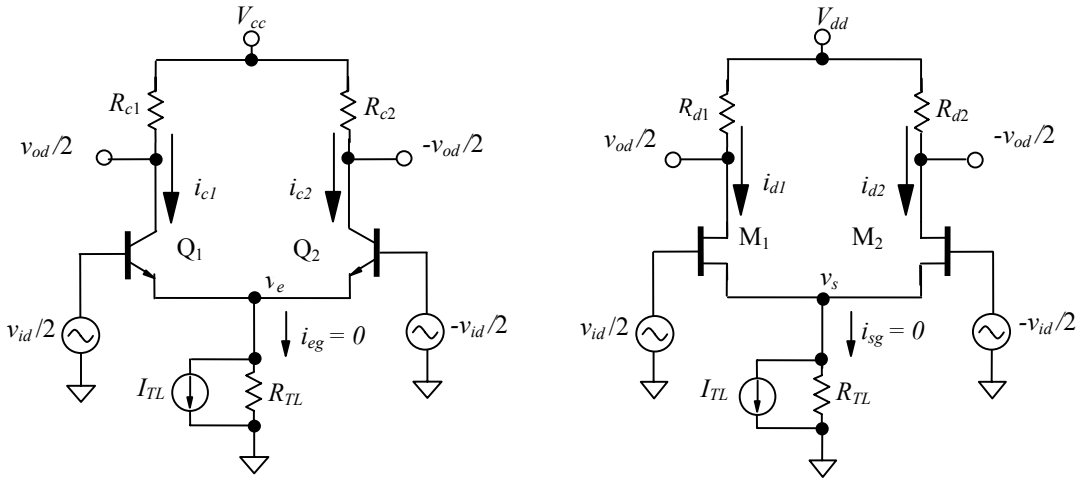
$$A_{cm} = \frac{A_{11} + A_{12} + A_{21} + A_{22}}{2} \quad , \tag{5.119}$$

$$A_{dm-cm} = \frac{A_{11} - A_{12} + A_{21} - A_{22}}{4} \quad , \tag{5.120}$$

$$A_{cm-dm} = A_{11} + A_{12} - A_{21} - A_{22} \quad , \tag{5.121}$$

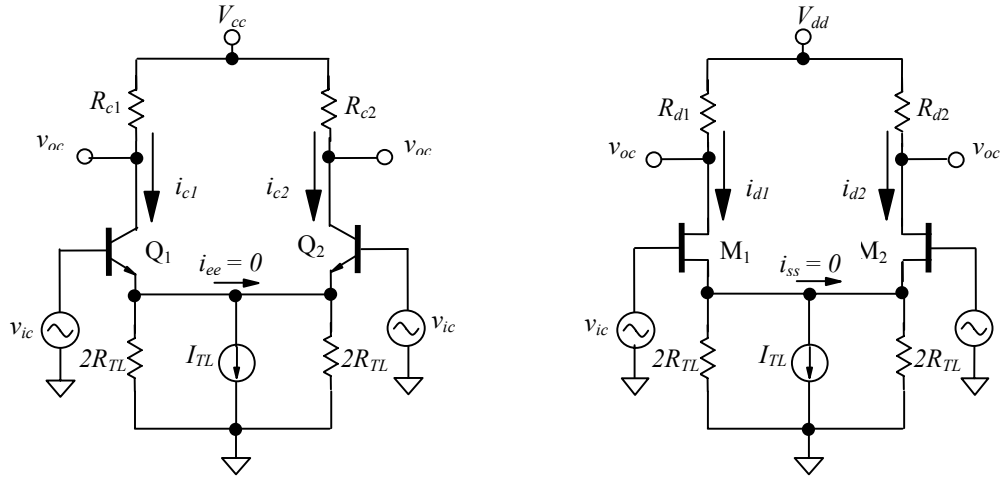
The calculation of $A_{dm\text{-}cm}$ and $A_{cm\text{-}dm}$ for bipolar transistor and MOSFET is quite tedious and we are not going to introduce the mathematical derivation here. Compared with A_{dm} and A_{cm}, they are usually a couple orders smaller and can be neglected. Let's confine our discussion to only A_{dm} and A_{cm}.

By means of equations (5.118) and (5.119), A_{dm} and A_{cm} can be calculated from A_{11}, A_{12}, A_{21}, and A_{22}. This is a somewhat complicated way to do so. One shortcut is to directly



(a) Bipolar devices (b) MOSFET devices

Figure 5.19 Differential mode portion of differential pairs



(a) Bipolar devices (b) MOSFET devices

Figure 5.20 Common mode portion of differential pairs

change the input and output voltages into the differential mode and common mode. From expressions (5.112) to (5.115), we have

$$v_{i1} = v_{ic} + \frac{v_{id}}{2} \quad , \tag{5.122}$$

$$v_{i2} = v_{ic} - \frac{v_{id}}{2} \quad , \tag{5.123}$$

$$v_{o1} = v_{oc} + \frac{v_{od}}{2} \quad , \tag{5.124}$$

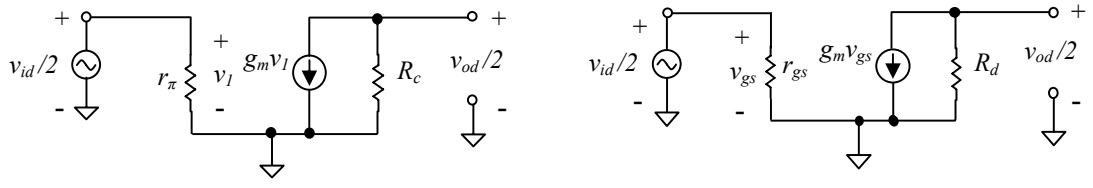
$$v_{o2} = v_{oc} - \frac{v_{od}}{2} \quad . \tag{5.125}$$

From expressions (5.122) to (5.125), we can find that a differential pair with input and output voltages, v_{i1}, v_{i2}, v_{o1}, v_{o2}, as shown in Figure 5.18 can be equivalent to the sum of two differential pairs as shown in Figure 5.19 and 5.20. One of them operates in differential mode with two inputs, $v_{id}/2$ and $-v_{id}/2$, and two outputs, $v_{od}/2$ and $-v_{od}/2$, and another operates in common mode with two inputs, v_{ic} and $-v_{ic}$, and two outputs, v_{oc} and $-v_{oc}$. In other words, the differential pair shown in Figure 5.18 can be replaced by Figures 5.19 and 5.20 for the differential mode and the common mode respectively. Theoretically this superposition principle is reasonable to the differential pair if only a small signal is concerned, in which the devices can be treated as linear components.

First let's discuss the case of a differential pair in differential mode as shown in Figure 5.19. Assuming that the differential pair has a perfect symmetry, the tail current i_{eg} or i_{sg} as shown in Figure 5.19, is

$$i_{eg} = i_{sg} = 0 \ , \tag{5.126}$$

because the currents contributing to i_{eg} from both devices Q1 and Q2 are the same in magnitude but the difference of 180^o in phase. Also, the currents contributing to i_{sg} from both devices M1 and M2 are the same in magnitude but the difference of 180^o in phase. The relation (5.126) implies that the emitters or sources of devices in Figure 5.19 are virtually grounded and the tail resistor R_{TL} is short-circuited in the AC or RF analysis. In other words, in either Figure 5.19 (a) or (b), the left and right branch of the differential pair are two identical and independent half circuit. The equivalent of differential mode is sketched in Figure 5.21 after a series of approximations are assumed.



(a) Bipolar device

Approximation:

$r_b \to 0$ $\quad C_\pi \to 0$

$r_o \to \infty$ $\quad C_{cs} \to 0$

$r_c \to 0$

(b) MOSFET device

Approximation:

$r_d \to \infty$ $\quad C_{gd} \to 0$

$C_{db} \to 0$ $\quad C_{gb} \to 0$

$C_{sb} \to 0$ $\quad C_{gs} \to 0$

Figure 5.21 Equivalent of differential mode half circuit

From Figure 5.21(a) we have

$$\frac{v_{od}}{2} = -g_m v_1 R_c = -g_m \frac{v_{id}}{2} R_c \quad , \tag{5.127}$$

From Figure 5.21(b) we have

$$\frac{v_{od}}{2} = -g_m v_{gs} R_d = -g_m \frac{v_{id}}{2} R_d \quad , \tag{5.128}$$

where

$$R_c = R_{c1} = R_{c2} \quad , \tag{5.129}$$

$$R_d = R_{d1} = R_{d2} \quad . \tag{5.130}$$

Two equations, (5.127) and (5.128), can be merged into one expression, such as

$$\frac{v_{od}}{2} = -g_m \frac{v_{id}}{2} R \quad , \tag{5.131}$$

if we denote R_c and R_d by R. Consequently,

$$A_{dm} = \frac{v_{od}}{v_{id}} = -g_m R \quad . \tag{5.132}$$

As a differential pair, the differential mode voltage gain A_{dm} is expected to be as high as possible. Obviously the devices with high g_m are preferred for a differential pair. On the other hand, the A_{dm} can be escalated by increasing R(R_c or R_d)'s value. However, it might be harmful to obtain a higher differential output power gain. Eventually, the effective means to enhance the A_{dm} is to increase the g_m of the devices.

Secondly, let's examine the case of a differential pair in common mode as shown in Figure 5.20. Assuming that the differential pair has a perfect symmetry, the current i_{ee} or i_{ss}, as shown in Figure 5.20, is

$$i_{ee} = i_{ss} = 0 \quad , \tag{5.133}$$

because the left and right branch of each differential pair are two identical, independent half circuits and each half is driven by the same input voltage, v_{ic}. The equivalent of the common mode half-circuit can be sketched after a series of approximations are assumed as shown in Figure 5.22.

As matter of fact, Figure 5.22 is a common-emitter or common-source amplifier with emitter or source degeneration. Its transconductance G_m can be calculated by means of simple KCL and KVL, though the calculation itself is somehow tedious. (Readers can refer to Paul Gary's book…chapter 4 for detailed mathematical derivation.)

For a differential pair build by bipolar devices,

$$G_m = \frac{i_o}{v_{ic}} = g_m \frac{g_m}{1 + 2 g_m R_{TL} \left(g_m r_\pi + 1 \right)} \quad , \tag{5.134}$$

For a differential pair built by MOSFETs,

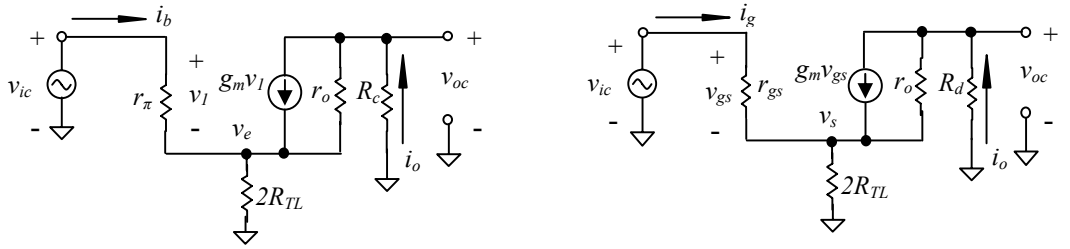
$$G_m = \frac{i_o}{v_{ic}} = \frac{g_m}{1 + 2g_m R_{TL}\left(g_m r_{gs} + 1\right)} \quad , \tag{5.135}$$

Let's denote r_π and r_{gs} by $r_{\pi gs}$. Consequently, (5.134) and (5.135) can be combined into one equation:

$$G_m = \frac{i_o}{v_{ic}} = \frac{g_m}{1 + 2g_m R_{TL}\left(g_m r_{\pi\, gs} + 1\right)} \quad , \tag{5.136}$$

when

$$r_o >> R_{TL} \ , \text{ or } \quad r_d >> R_{TL} \, . \tag{5.137}$$



(a) Bipolar device

Approximation:

$r_b \to 0$ $\quad C_\pi \to 0$

$r_c \to 0$ $\quad C_{cs} \to 0$

(b) MOSFET device

Approximation:

$C_{db} \to 0$ $\quad C_{gd} \to 0$

$C_{sb} \to 0$ $\quad C_{gb} \to 0$

$C_{gs} \to 0$

Figure 5.22 Equivalent of common mode half circuit

On the other hand,

$$v_{oc} = -i_o R = -G_m R v_{ic} \quad , \tag{5.138}$$

then

$$A_{cm} = \frac{v_{oc}}{v_{ic}} = -G_m R = -\frac{g_m R}{1 + 2g_m R_{TL}\left(g_m r_{\pi gs} + 1\right)} \quad . \tag{5.139}$$

As a differential pair, the common mode voltage gain A_{cm} is expected to be as low as possible because usually the purpose of a differential pair is to magnify the differential signal while its common mode signal must be suppressed as much as possible. The common mode operation in a differential pair produces DC offset, which is very harmful to the RF signal in a direct conversion communication system.

From equation (5.139) it can be seen that the tail resistor R_{TL} is a key part in suppressing the common mode voltage gain. Should R_{TL} is zero, the A_{cm} will be equal to A_{dm} as shown in equation (5.132) and the differential pair loses its special meaning.

In order to quantitatively measure the merit of magnifying the differential mode but suppressing the common mode signal, a special parameter, CMRR (Common Mode Rejection Ratio) is defined as the ratio of the differential mode voltage gain to the common mode voltage gain, that is

$$CMRR = \frac{A_{dm}}{A_{cm}} = 1 + 2g_m R_{TL}\left(1 + \frac{1}{g_m r_{\pi gs}}\right) \quad . \tag{5.140}$$

It measures the rejection capability opposed the common mode voltage and the magnifying capability to the differential mode voltage. A good differential pair must have high A_{dm} but low A_{cm}. From equation (5.140) it can be seen that a higher g_m of the devices and a higher tail resistance R_{TL} are preferred.

It is meaningless to talk about CMRR for a single-ended block because the "differential," or 180^o phase difference, does not exist.

A pseudo-differential pair is shown in Figure 5.23. It consists of two identical and independent amplifiers. Figure 5.23(a) shows a pseudo-differential pair without a tail or degeneration resistor and Figure 5.23(b) shows a pseudo-differential pair without a tail resistor but with two separate emitter or source degeneration resistors R_{e1} and R_{e2} or R_{s1} and R_{s2}. A tail resistor must be a common coupling resistor for both Q_1 and Q_2 or M_1 and M_2.

What would the value of CMRR be in a pseudo-differential pair?

In both cases shown in Figure 5.23, there is no coupling resistor R_{TL}, that is

$$R_{TL} = 0 \ , \tag{5.141}$$

From expression (5.140), the CMRR of these two pseudo-differential pairs is

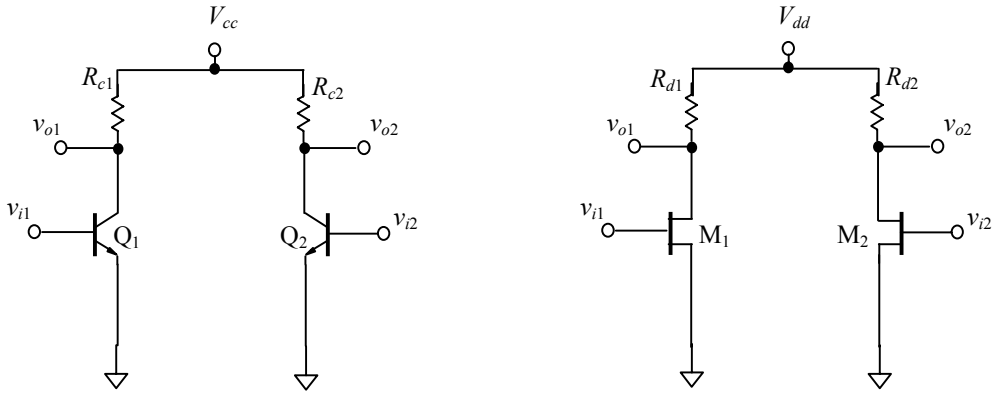
$$CMRR = 1 \ , \tag{5.142}$$

In the logarithmic scale, the expression (5.142) becomes

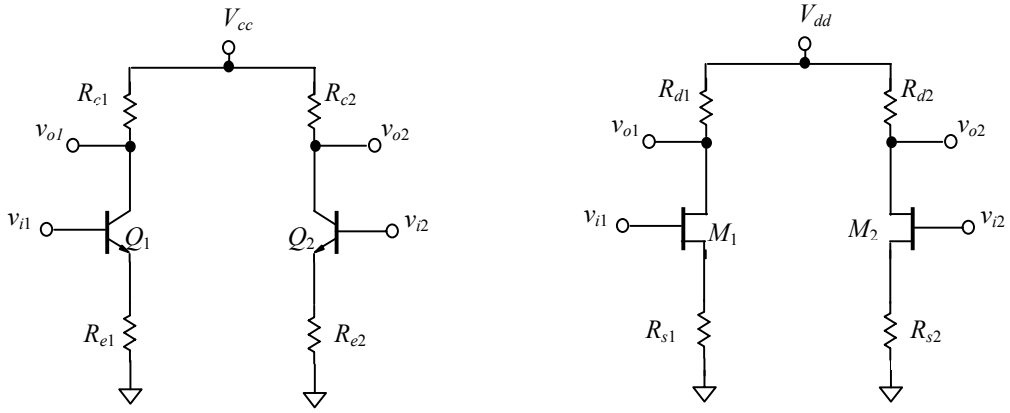
$$CMRR = 0 \ \text{dB}. \tag{5.143}$$

It is therefore concluded that there is no any common mode rejection capability in a pseudo-differential pair. Comparing Figure 5.23 with Figure 5.19, it can be found that in a pseudo-differential pair, there is no common coupling part, resistor R_{TL}, which

combines two individual single-ended blocks together. This is the reason for *CMRR*=0 dB as shown in expression (5.143) and named as "pseudo-differential pair".



(a) R_{TL}=0, R_{e1}=0, R_{e2}=0
R_{TL}=0, R_{s1}=0, R_{s2}=0



(b) R_{TL} =0, R_{e1}≠0, R_{e2}≠0
R_{TL} =0, R_{s1}≠0, R_{s2}≠0

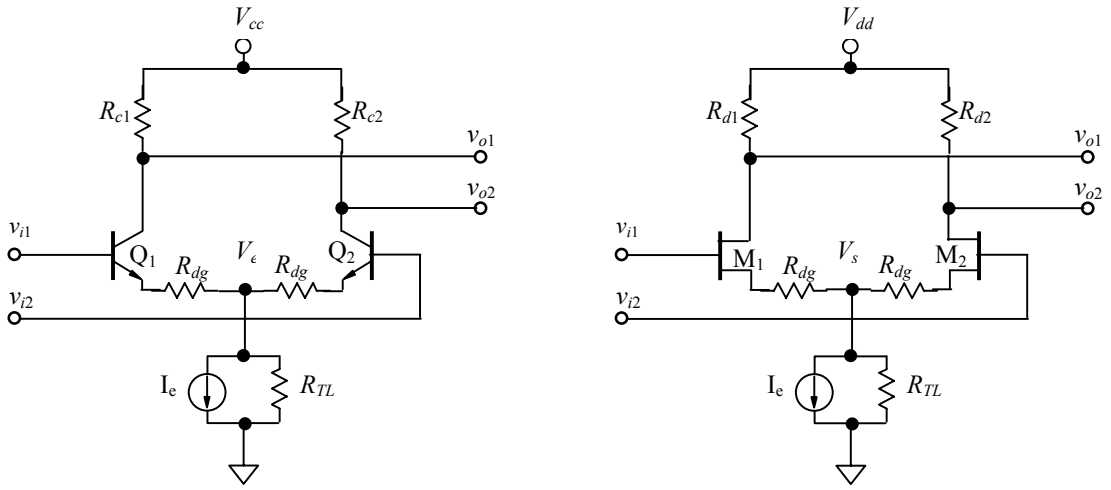
Figure 5.23 Pseudo-differential pair.

5.2.3 Improvement of CMRR

The CMRR can be improved by inserting a degeneration resistor R_{dg} between each emitter or source and R_{TL}, as shown in Figure 5.24.

Equations (5.132), (5.139), and (5.140) are thus modified to

$$A_{dm} = \frac{-g_m R}{1 + g_m R_{dg}\left(1 + \dfrac{1}{g_m R_{\pi gs}}\right)} \quad , \tag{5.144}$$



(a) Bipolar device (b) MOSFET device

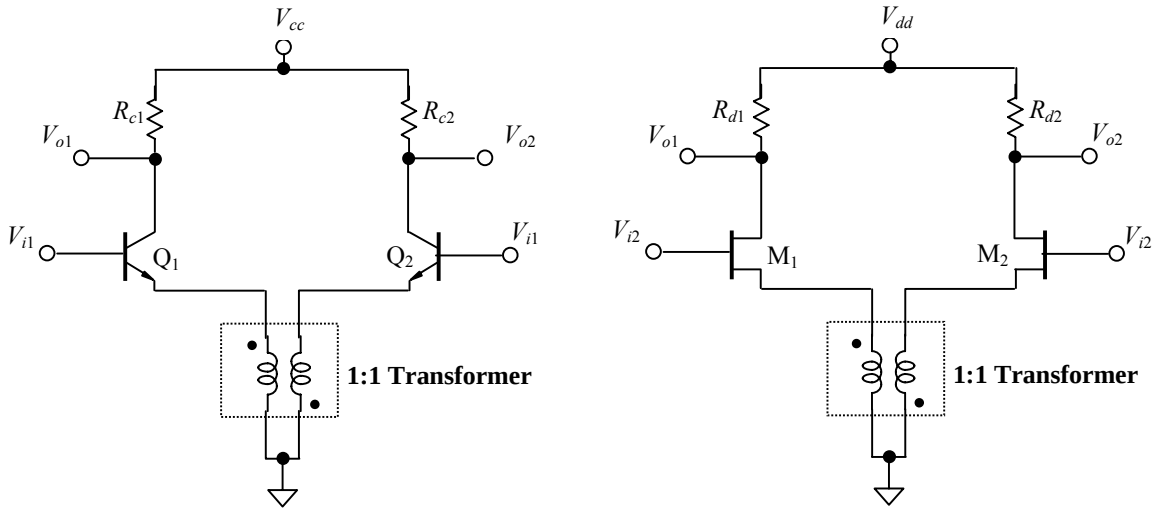
Figure 5.24 Emitter or source coupled pair with degeneration resistors

$$A_{cm} \approx \frac{-g_m R}{1 + 2g_m R_{TL}\left(1 + \dfrac{1}{g_m R_{\pi gs}}\right)\left(1 + \dfrac{R_{dg}}{2R_{TL}}\right)} \quad , \tag{5.145}$$

$$CMRR \approx \frac{A_{dm}}{A_{cm}} = \frac{1 + 2g_m R_{TL}\left(1 + \dfrac{1}{g_m R_{\pi gs}}\right)\left(1 + \dfrac{R_{dg}}{2R_{TL}}\right)}{1 + g_m R_{dg}\left(1 + \dfrac{1}{g_m R_{\pi gs}}\right)} \quad . \tag{5.146}$$

Another scheme to improve CMRR is to insert a transformer between the emitters or sources and the ground as shown in Figure 5.25. This is a simple but effective idea. The transformer not only plays the part of coupling but also forces the source current in each branch to have the same magnitude with 180^o phase difference. It prevents the production of common mode output voltage and therefore improves the CMRR.

In addition, the transformer also improves the linearity of the differential pair.



(a) Bipolar device (b) MOSFET device

Figure 5.25 Improvement of CMRR by a 1:1 transformer

It should be noted, however, that the improvement of CMRR depends considerably on how exactly symmetrical the transformer is. An asymmetrical transformer might even result in a degradation of the CMRR performance.

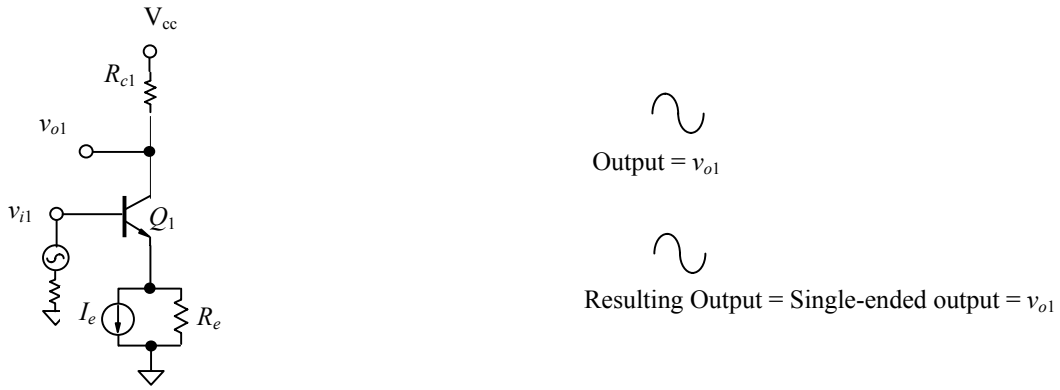
5.2.4 Increase of Voltage Swing

Figure 4.25 shows a comparison of the voltage swing between a single ended block and a differential pair. The output of a differential pair is about double the output of a single ended block if

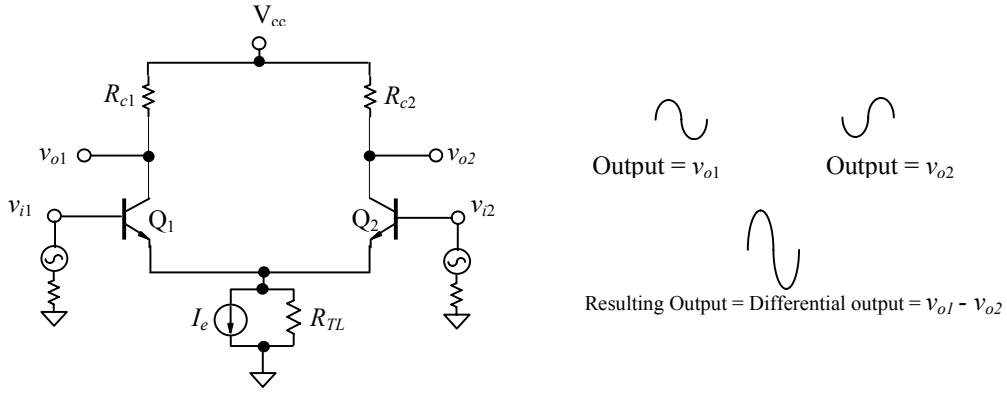
$$|v_{i1}| = |v_{i2}| \quad , \tag{5.147}$$

$$\angle v_{i1} - \angle v_{i2} = 180^o \quad . \tag{5.148}$$

The apparent reason is due to the differential configuration. The price of the double voltage swing obtained in a differential pair, however, is a double current drain, more noise, and a two-fold IC die area opposed to a single-ended block.



(a) A single-ended stage.



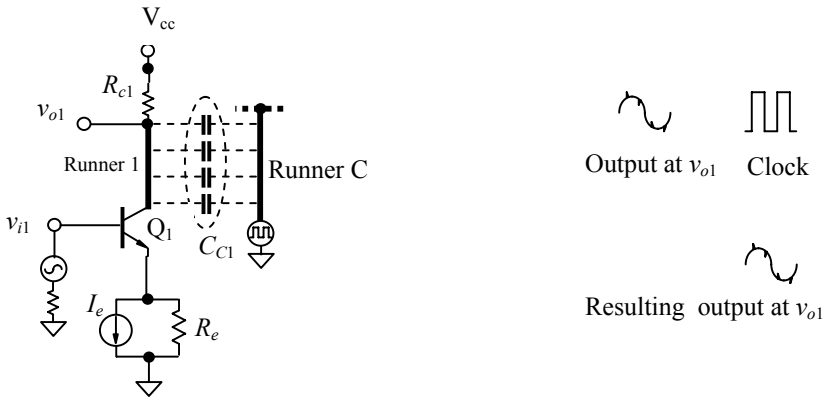
(b) A differential pair.

Figure 5.26 Increase of the voltage swing in a differential pair

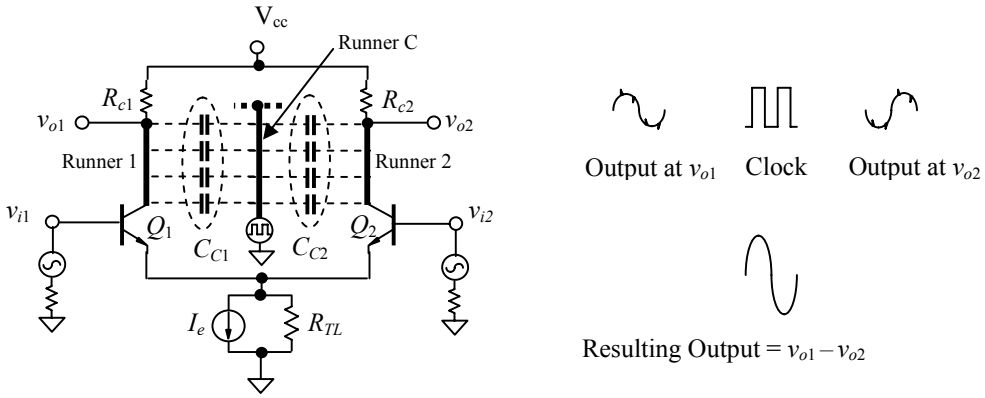
5.2.5 Cancellation of Interference

One of the advantages for a differential pair is the capability of interference cancellation. Figure 5.27 shows this capability of rejection.

Figure 5.27 (a) shows the interference on a single-ended block. Runner 1 is an actual runner from collector to resistor R_{c1} in the layout of a single-ended block. Assuming that runner 1 is in parallel with and near to runner C, on which a clock signal is carried. The clock frequency is double that of the input signal frequency, and the interference appears due to the spray capacitors C_{c1} between runner 1 and runner C. The clock transit pulses are superimposed on the output signal as up- and down-ward spikes. The output signal v_{o1} is therefore interfered by the clock signal.



(a) A single-ended stage.



(b) A differential pair.

Figure 5.27 Cancellation of interference in a differential pair

Now let's move on to Figure 5.27 (b), which shows the interference in a differential pair. Assuming that runner C is exactly parallel between runners 1 and 2 and the differential pair has a perfect symmetrical layout, the spikes representing the interference on the two output signals have the same polarity while the output signals themselves have 180^{o} phase difference. Consequently, the interferences at the resulting differential output are cancelled from each other. Figure 5.27(b) demonstrates such a cancellation of interference by their corresponding waveforms.

Obviously, the cancellation of interferences becomes realistic only if the layout of the differential pair has a perfect symmetrical status and if the source of the interferences are located exactly at the middle line between runners 1 and 2, or at the symmetrical axis of the differential pair layout. Of course, interferences can still be rejected partially if the differential pair layout is not in a perfect symmetrical condition.

Such a cancellation of interference has sometimes been inferred as a kind of "rejection of common mode noise" by the differential pair. This is an inappropriate assertion. First, the clock pulses as shown in Figure 5.27 are not noise but are regular and definite signals. Secondly, noise is a random, stochastic process so that it is impossible to distinguish it into "common mode noise" and "differential mode noise." Should it be possible to do that, all the engineers over the world would be very happy because at least, the "common mode noise" could be rejected by a differential pair.

5.2.6 Noise in a Differential Pair

Is it possible to cancel or partially cancel out the noise in a differential pair? There are different opinions over such a controversy.

Figures 5.28 shows a noise source from the DC power supply and Figure 5.29 shows another noise source from the tail transistor Q_o. They are good examples to answer the question.

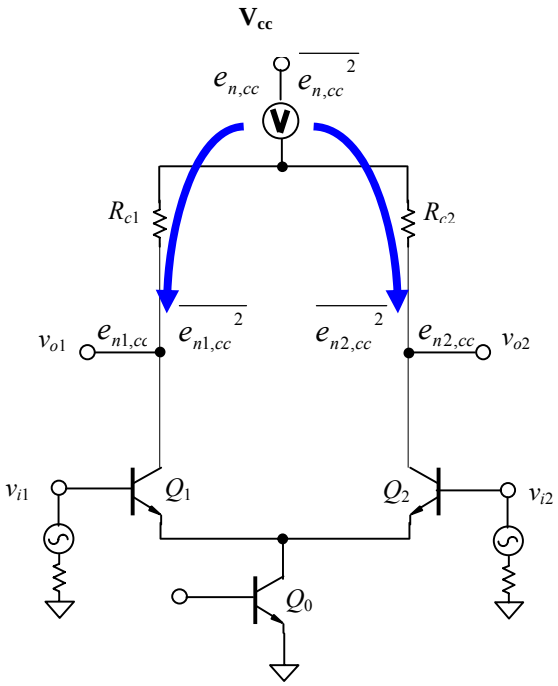


Figure 5.28 At differential outputs, noise from DC power supply cannot be "differentiated away"!

Let's denote

$e_{n,cc}$ = Noise voltage of noise source from the DC power supply, a random variable,

$\overline{e_{n,cc}^2}$ = Square of noise voltage's variance of noise source from the DC power supply,

$e_{n,Qo}$ = Noise voltage of noise source from the tail transistor Q_o, a random variable,

$\overline{e_{n,Qo}^{\,2}}$ = Square of noise voltage's variance of noise source from the tail transistor Q_o.

In the case of the noise source from the DC power supply as shown in Figure 5.28, the noise voltages and their square of its variances can be found at two output terminals :

At output v_{o1} : the noise voltage is $e_{n1,cc}$ and its square of variance is $\overline{e_{n1,cc}^{\,2}}$;

At output v_{o2} : the noise voltage is $e_{n1,cc}$ and its square of variance is $\overline{e_{n1,cc}^{\,2}}$.

The noises at these two terminals should be partially or fully correlated as long as

$$R_{c1} = R_{c2} \ , \tag{5.149}$$

and the layout of the differential pair is perfectly symmetrical. The main reason is that these two noises coming from the same noise source at the DC power supply. It is therefore inferred that

$$\overline{e_{n1,cc}^{\,2}} = \overline{e_{n2,cc}^{\,2}} \ , \tag{5.150}$$

and

$$e_{n1,cc} = e_{n2,cc} \ . \tag{5.151}$$

Consequently, on the basis of the expressions (5.150) and (5.151) it is concluded that the noise from the DC power supply can be cancelled by a differential pair if the differrential pair is perfectly symmetrical.

In the case of the noise source from the tail transistor Q_o as shown in Figure 5.29, the noise voltages and their square of its variances can be found at two output terminals :

At output v_{o1} : the noise voltage is $e_{n1,Qo}$ and its square of variance is $\overline{e_{n1,Qo}^{\,2}}$;

At output v_{o2} : the noise voltage is $e_{n1,Qo}$ and its square of variance is $\overline{e_{n1,Qo}^{\,2}}$.

The noises at these two terminals should be partially or fully correlated as long as two transistors Q1 and Q2 are perfectly identical, that is

$$Q_1 = Q_2 \ , \tag{5.152}$$

and the layout of the differential pair is perfectly symmetrical. The main reason is that these two noises coming from the same noise source at the tail transistor Q_o. It is therefore inferred that

$$\overline{e_{n1,Qo}^{\,2}} = \overline{e_{n2,Qo}^{\,2}} \ , \tag{5.153}$$

and

$$e_{n1,Qo} = e_{n2,Qo} \ . \tag{5.154}$$

Consequently, on the basis of the expressions (5.153) and (5.154) it is concluded that the noise from the tail transistor Q_o can be cancelled by a differential pair if the differrential pair is perfectly symmetrical.

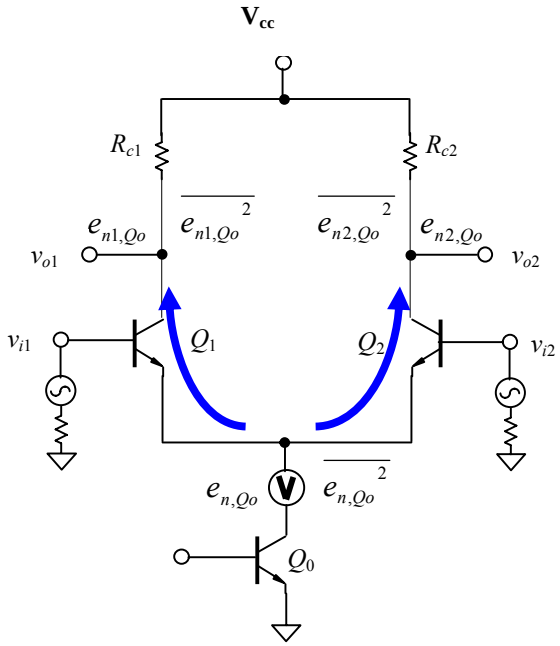


Figure 5.29 At differential outputs, noise from Q_0 cannot be "differentiated away"!

From the viewpoint of noise reduction, it seems very encouraged to replace the single-ended stage by a differential pair if the inferences above are true. As a matter of fact, the expressions (5.150) and (5.153) are convincible because the average noise powers at two output terminals would be the same as long as the differential pair is perfectly symmetrical.

However, the expressions (5.151) and (5.154) might not be possible because the noise voltages is a random variable. In a resistor or even in a perfect block conductor, the motion of electrons or hole carriers is a random, stochastic process. Their interference with, attenuation or magnification of the incoming noise from the DC power supply or from the tail transistor Q_o is, thus, also a random, stochastic process. At any instant, it is impossible to keep these two noise voltages, $e_{n1,cc}$ and $e_{n2,cc}$, or $e_{n1,Qo}$ and $e_{n2,Qo}$, in equal magnitude and in equal phase. On other words, at any instant,

$$e_{n1,cc} \neq e_{n2,cc}, \tag{5.155}$$

and

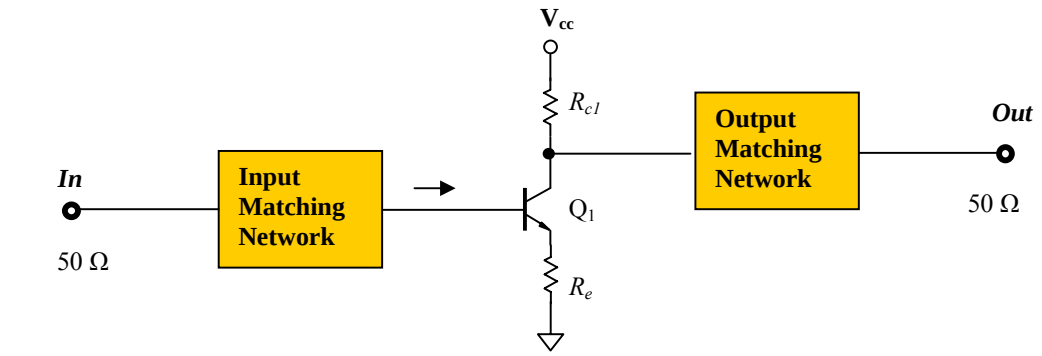
$$e_{n1,Qo} \neq e_{n2,Qo}. \tag{5.156}$$

It is therefore inferred that the noise figure of a differential pair is higher than that of the single-ended stage if the differential pair is presumably constructed by two single-ended branches.

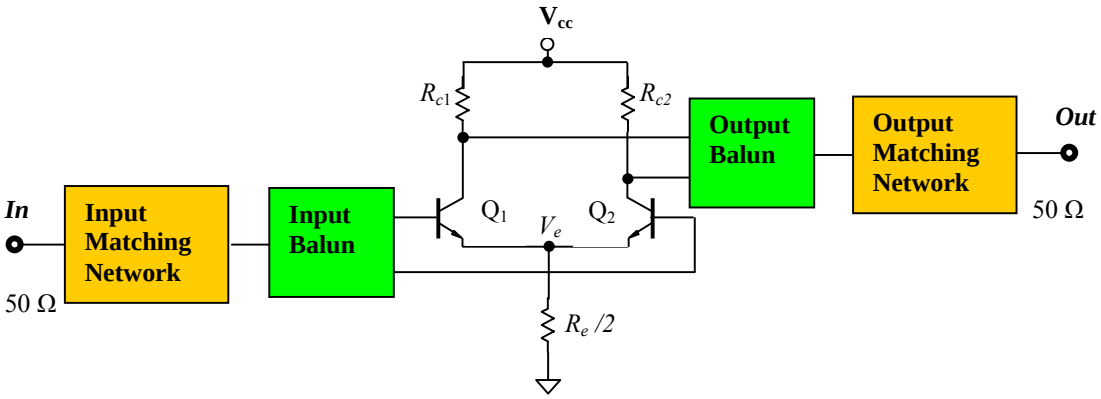
There might be another questions:

Is it possible that the noise can be partially differenciated-out by a differential pair if the noise voltages at two output terminals are partially, though not totally, correlated?

Is it possible that the noise figure in a differential pair less than a single-ended block if the former is built basically by two single-ended blocks?



(a) A single-ended stage



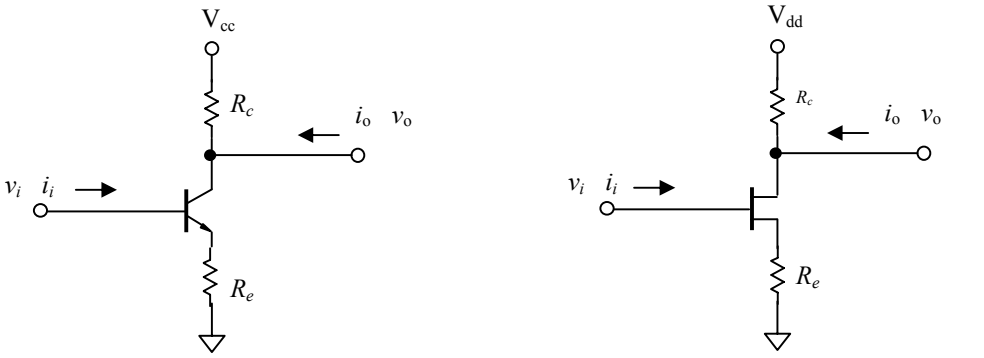
(b) A differential pair

Figure 5. 30 Experimentally test circuits for a comparison of noise figures between single-ended and differential pair.

The correctness of a theory can be verified by experiment. It is therefore suggested to test the noise figures for a single-ended stage and a differential pair as shown in Figure 5.30. The differental pair is presumably constructed by two single-ended branches. Of course, both of them must be impedance-matched by the impedance matching networks, and, the effects of baluns and matching networks on the noise figure must be calibrated. Finally, the current drain for the differential pair is approximately double of that in the single-ended stage.

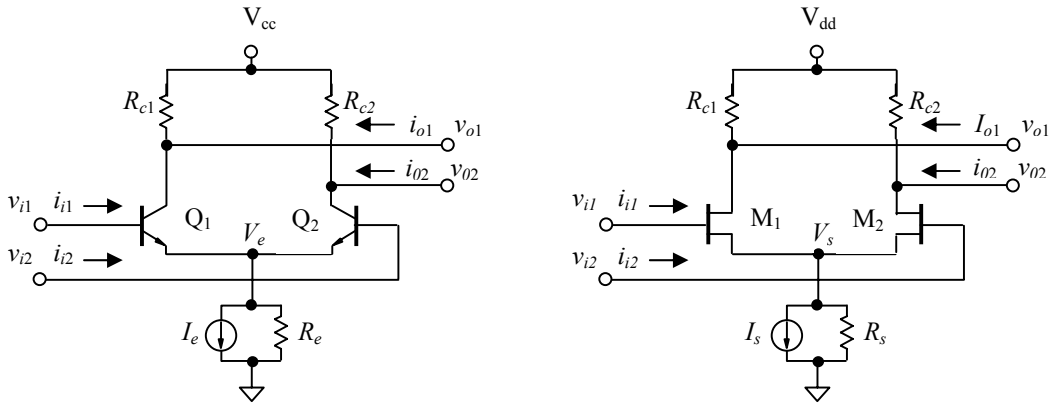
5.3 Apparent Difference between Single-ended Stage and Differential Pair

In the previous two sections we have discussed the single-ended block and differential pair individually. Now, we are going to briefly summarize and compare their differences.



(a) A bipolar single-ended stage (b) A MOSFET single-ended stage

Figure 5. 31 Single-ended stages.



(a) A bipolar differential pair (b) A MOSFET differential pair

Figure 5. 32 Differential pairs.

In most RF blocks, such as the LNA, mixer, VCO, filter, and PA, both the single-ended block and differential pair has been developed. Figure 5.31 depicts a typical single-ended block and Figure 5.32 depicts a typical differential pair: an emitter coupled pair. In the history of the electronic circuit development, the single-ended block was studied and

implemented in the early stages. Gradually people found that some special types of performance could be only be obtained by a differential pair instead of a single-ended stage, such as a push-pull amplifier or a double-balanced mixer. In recent years increasing numbers of differential circuits have appeared in RF or RFIC designs, especially in the "zero" IF or direct conversion communication systems.

Just as its name implies, the single-ended circuit has only one input port and one output port. Both input and output ports are referenced to the ground. The grounding port is referenced as an RF zero potential. In the differential pair, there are two input and output ports respectively. Either at input or output differential ports, the signals have same magnitude but 180^o of phase difference.

For a differential pair, the input and output signals can be expressed by voltage as shown below:

$$v_{i1} = -v_{i2} \quad , \tag{5.157}$$

or

$$|v_{i1}| = |v_{i2}| \quad , \quad \angle v_{i1} - \angle v_{i2} = 180^o , \tag{5.158}$$

and,

$$v_{o1} = -v_{o2} \quad , \tag{5.159}$$

or

$$|v_{o1}| = |v_{o2}| \quad , \quad \angle v_{o1} - \angle v_{o2} = 180^o \tag{5.160}$$

Instead of voltage, the input and output signals can be expressed by current as well:

$$i_{i1} = -i_{i2} \quad , \tag{5.161}$$

or

$$|i_{i1}| = |i_{i2}| \quad , \quad \angle i_{i1} - \angle i_{i2} = 180^o , \tag{5.162}$$

and,

$$i_{o1} = -i_{o2} \quad , \tag{5.163}$$

or

$$|i_{o1}| = |i_{o2}| \quad , \quad \angle i_{o1} - \angle i_{o2} = 180^o . \tag{5.164}$$

Sometimes the two ports of the differential pair are marked with positive and negative signs respectively. It does not mean that the potential of one port is always positive and that of another is always negative whereas it does mean that they are always kept a 180^o of phase difference alternatively.

The number of parts of a differential pair is about double that of the single-ended circuit. Consequently, either in the RF circuit with discrete parts or in the RFIC design, the layout area for the differential pair is about double than that of the single-ended circuit.

If a differential pair is built basically by two single-ended blocks, its current drain is usually double of that in a single-ended stage and its noise figure is higher than a single-ended stage.

A differential input signal is expected to be faithfully magnified or transported by a differential pair. Assuming that the input of the differential pair is a pure differential signal, the output of the differential pair thus should be a pure differential signal as well. In order to approach the ideal differential performance, the symmetry of the layout must be maintained as much as possible. For a single-ended circuit, symmetry is meaningless.

In addtion, a special part called balun is needed for the differential pair. The "balun" is a new word coined by combining the words "**bal**anced" and "**un**balanced." A balun converts a differential signal into two single-ended signals, or vice versa.

Table 5.3 Difference of appearance between single-ended stage and differential pair.

	Single-ended stage	Differential pair
Parts	1x	≈2x
Layout area	1x	≈2x
Current drain	1x	2x
Noise figure	1x	>1x
Symmetry	No	Yes
Balun needed	No	Yes, if transformation between single-ended stage and differential pair is needed.

Table 5.3 summarizes the primary comparison as described above. From such a comparison between the single-ended block and the differential pair, it seems that the differential pair is devoid of any merit. If so, what is the differential pair for?

The answer will be found in next section.

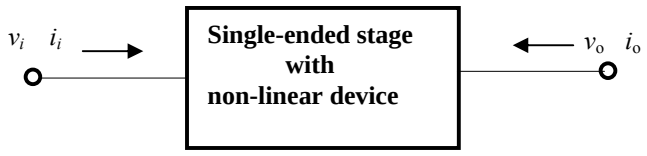
5.4 DC Offset

All the items listed in Table 5.3 show the defects of the differential pair in respect to the single-ended stage. However, there is an outstanding feature of the differential pair which is not listed in Table 5.3. This feature is the dramatic reduction of the so-called DC offset.

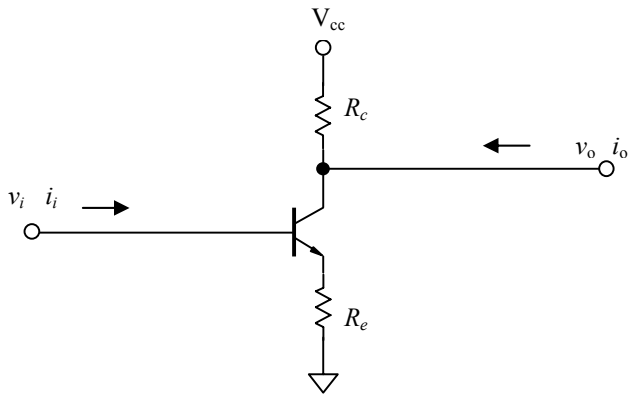
DC offset comes from the non-linearity of devices and an imperfect symmetrical layout of the circuit block. Theoretically, DC offset exists in a single-ended block but not in an ideal differential pair.

In a "zero" IF or direct conversion communication system, the cancellation of DC offset is of extreme importance. This is why the differential pair is widely applied in circuitry today regardless of all the defects listed in Table 5.3.

5.4.1 DC Offset in a Single-ended Device



(a) Block diagram



(b) Schematic

Figure 5.33 A single-ended stage with non-linear device

It is well known that non-linearity exists in all devices, such as diodes, varactors, bipolar transistors, MOSFET's, and so on. In general, the non-linearity of a device includes both odd and even orders from 0, 1, 2, 3, to infinity. A single-ended block built by devices with non-linearity should have non-linearity in a similar mathematical form, that is, it also includes both odd and even orders from 0, 1, 2, 3, to infinity.

Similar to a single device with non-linearity, if the input signal in Figure 5.33 is sinusoidal, that is,

$$v_i = v_{io}\cos\omega t \quad , \tag{5.165}$$

where v_i = Sinusoidal input signal,
v_{io} = Amplitude of the input signal,
ω = Angular frequency of the sinusoidal signal.

Then the transfer function of a single-ended stage can be expressed by

$$\begin{aligned}
i_o &= a_o + a_1 v_i + a_2 v_i^2 + a_3 v_i^3 + a_4 v_i^4 + \cdots\cdots \\
&= a_o + a_1 v_{io}\cos\omega t + a_2 v_{io}^2 \cos^2\omega t + a_3 v_{io}^3 \cos^3\omega t + a_4 v_{io}^4 \cos^4\omega t + \cdots\cdots \\
&= a_o + a_1 v_{io}\cos\omega t + a_2 v_{io}^2[1/2 + (1/2)\cos 2\omega t] + a_3 v_{io}^3[(3/4)\cos\omega t + (1/4)\cos 3\omega t] \\
&\quad + a_4 v_{io}^4[3/8 + (4/8)\cos 2\omega t + (1/8)\cos 4\omega t] + \cdots\cdots \\
&= [a_o + (1/2)a_2 v_{io}^2 + (3/8)\, a_4 v_{io}^4 + \cdots\cdots] \\
&\quad + [a_1 v_{io} + (3/4)\, a_3 v_{io}^3 + \cdots\cdots]\cos\omega t \\
&\quad + [(1/2)a_2 v_{io}^2 + (4/8)\, a_4 v_{io}^4 + \cdots\cdots]\cos 2\omega t \\
&\quad + [(1/4)a_3 v_{io}^3 + \cdots\cdots]\cos 3\omega t \\
&\quad + [(1/8)\, a_4 v_{io}^4 + \cdots\cdots]\cos 4\omega t + \cdots\cdots
\end{aligned} \tag{5.166}$$

where a_i = ith order of coefficient of non-linearity.

The DC current containing in equation (5.166) consists of two parts, $i_{o,DC}$ and DC offset, and is denoted by $i_{o,dc}$. They can be expressed by

$$\begin{aligned}
i_{o,\,dc} &= i_{o,DC} + \text{DC offset} \\
&= a_o + (1/2)a_2 v_{io}^2 + (3/8)\, a_4 v_{io}^4 + \cdots\cdots
\end{aligned} \tag{5.167}$$

$$i_{o,\,DC} = a_o \ , \tag{5.168}$$

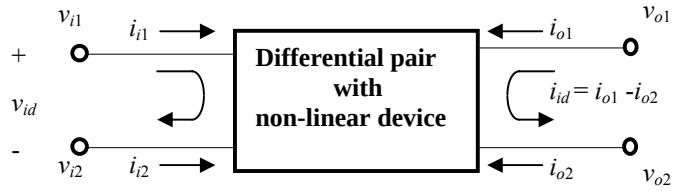
$$\text{DC offset} = (1/2)a_2 v_{io}^2 + (3/8)\, a_4 v_{io}^4 + \cdots\cdots \tag{5.169}$$

The 1st part of DC output current, $i_{o,DC}$, called "intrinsic" DC output current, is independent from the input signal. The 2nd part of DC output current, named "DC offset" output current, is due to the non-linearity of the block and relies directly on the RF input signal. If the input is a DC voltage, the total DC output current only consists of $i_{o,DC}$ part while the DC offset part disappears.

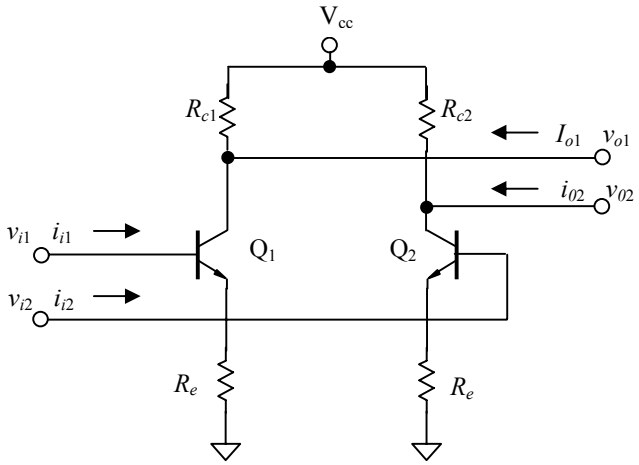
A remarkable feature of DC offset is that it is contributed by only the even orders but not the odd orders of non-linearity.

5.4.2 Zero DC Offset in a Pseudo-Differential Pair

A differential pair without coupling as shown in Figure 5.34 is a combined entity by two independent but identical single-ended blocks. Let's call it a pseudo-differential pair.



(a) Block diagram



(b) schematic

Figure 5.34 A differential pair with non-linear devices without coupling

Theoretically this pseudo-differential pair can operate like a real differential pair as long as we place the input differential signal to the inputs of the two single-ended stages and take the output differential signal from the outputs of the two single-ended blocks. In other words, the input of the 1st single-ended stage, the input of the 2nd stage, the output of the 1st stage, and the output of the 2nd stage, are treated as v_{i1}, v_{i2}, v_{o1}, and v_{o2}, respectively.

The essential difference between a pseudo- and a real differential pair is the capability of CMRR. In a practical design, a real differential pair usually has a pretty high CMRR capability whereas a pseudo-differential pair cannot reject any common mode interference. However, the DC offset in a real differential pair will be the same as in a pseudo-differential pair. The analysis of DC offset for a pseudo-differential pair is much easier than for a real differential pair. By means of an analysis of DC offset in a pseudo-differential pair, it is enable to provide a clear picture how the DC offset can be significantly reduced in a differential pair but not in a single-ended stage. Therefore, let's analyze the DC offset reduction in a pseudo-differential pair.

Assuming that the input signals in Figure 5.34 are of differential sinusoidal voltages,

$$v_{i1} = v_{io}\cos\omega t \quad , \tag{5.170}$$

$$v_{i2} = v_{io}\cos(\omega t + 180^o) = -v_{io}\cos\omega t \quad , \tag{5.171}$$

where v_{i1}, v_{i2} = Sinusoidal input signals,
v_{io1}, v_{io2} = Amplitude of the input signals,
ω = Angular frequency of the sinusoidal signal.

Similar to a single device with non-linearity, the transfer function of each single-ended device can be expressed by

$$\begin{aligned} i_{o1} &= a_o + a_1 v_{i1} + a_2 v_{i1}^2 + a_3 v_{i1}^3 + a_4 v_{i1}^4 + \cdots\cdots \\ &= a_o + a_1 v_{io}\cos\omega t + a_2 v_{io}^2 \cos^2\omega t + a_3 v_{io}^3 \cos^3\omega t + a_4 v_{io}^4 \cos^4\omega t + \cdots\cdots \\ &= a_o + a_1 v_{io}\cos\omega t + a_2 v_{io}^2[1/2 + (1/2)\cos 2\omega t] + a_3 v_{io}^3[(3/4)\cos\omega t + (1/4)\cos 3\omega t] \\ &\quad + a_4 v_{io}^4[3/8 + (4/8)\cos 2\omega t + (1/8)\cos 4\omega t] + \cdots\cdots \end{aligned}$$

$$\begin{aligned} i_{o1} = {} & [a_o + (1/2)a_2 v_{io}^2 + (3/8)\, a_4 v_{io}^4 + \cdots\cdots] \\ & + [a_1 v_{io} + (3/4)\, a_3 v_{io}^3 + \cdots\cdots \quad]\cos\omega t \\ & + [(1/2)a_2 v_{io}^2 + (4/8)\, a_4 v_{io}^4 + \cdots\cdots \quad]\cos 2\omega t \\ & + [(1/4)a_3 v_{io}^3 + \cdots\cdots \quad]\cos 3\omega t \\ & + [(1/8)\, a_4 v_{io}^4 + \cdots\cdots \quad]\cos 4\omega t + \cdots\cdots \end{aligned} \tag{5.172}$$

$$\begin{aligned} i_{o2} &= a_o + a_1 v_{i2} + a_2 v_{i2}^2 + a_3 v_{i2}^3 + a_4 v_{i2}^4 + \cdots\cdots \\ &= a_o - a_1 v_{io}\cos\omega t + a_2 v_{io}^2 \cos^2\omega t - a_3 v_{io}^3 \cos^3\omega t + a_4 v_{io}^4 \cos^4\omega t - \cdots\cdots \\ &= a_o - a_1 v_{io}\cos\omega t + a_2 v_{io}^2[1/2 + (1/2)\cos 2\omega t] - a_3 v_{io}^3[(3/4)\cos\omega t + (1/4)\cos 3\omega t] \\ &\quad + a_4 v_{io}^4[3/8 + (4/8)\cos 2\omega t + (1/8)\cos 4\omega t] - \cdots\cdots \\ &= [a_o + (1/2)a_2 v_{io}^2 + (3/8)\, a_4 v_{io}^4 + \cdots\cdots] \\ &\quad + [-a_1 v_{io} - (3/4)\, a_3 v_{io}^3 + \cdots\cdots \quad]\cos\omega t \end{aligned}$$

$$+ [(1/2)a_2 v_{io}^2 + (4/8)\, a_4 v_{io}^4 + \ldots\ldots \quad]\cos 2\omega t$$
$$+ [-(1/4)a_3 v_{io}^3 + \ldots\ldots \quad]\cos 3\omega t$$
$$+ [(1/8)\, a_4 v_{io}^4 + \ldots\ldots \quad]\cos 4\omega t + \ldots\ldots \tag{5.173}$$

Two assumptions have been applied in the derivation of expressions (5.172) and (5.173). First, the differential pair is assumed to be in a perfect symmetrical status so that the non-linearity coefficients in the calculation of i_{o1} and i_{o2} devices are exactly the same. Secondly, the assumed input equations (5.170) and (5.171) have been applied in the derivation of expressions (5.172) and (5.173).

Thus, the output differential current is

$$i_o = i_{o1} - i_{o2} = 2[a_1 v_{io} + (3/4)\, a_3 v_{io}^3 + \ldots\ldots]\cos\omega t + 2[(1/4)a_3 v_{io}^3 + \ldots\ldots]\cos 3\omega t + \ldots\ldots \tag{5.174}$$

It is therefore concluded that, with a perfect differential pair, both the intrinsic DC output current $i_{o,DC}$ and the DC offset current due to non-linearity are removed.

From expressions (5.172) and (5.173) one can find the output DC currents are

$$i_{o1,\,dc} = i_{o2,\,dc} = a_o + (1/2)a_2 v_{io}^2 + (3/8)\, a_4 v_{io}^4 + \ldots\ldots , \tag{5.175}$$

therefore,

$$i_{o,dc} = i_{o1,\,dc} - i_{o2,\,dc} = 0 , \tag{5.176}$$

then,

$$i_{o,\,DC} = 0 , \tag{5.177}$$

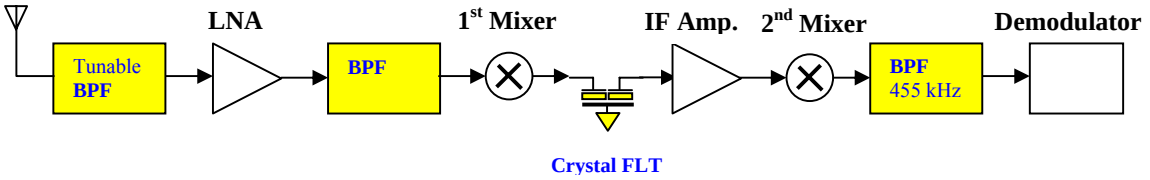
$$\text{DC offset} = 0 . \tag{5.178}$$

The even orders of non-linearity are theoretically cancelled from each other in an ideal differential configuration, thus resulting in zero DC offset.

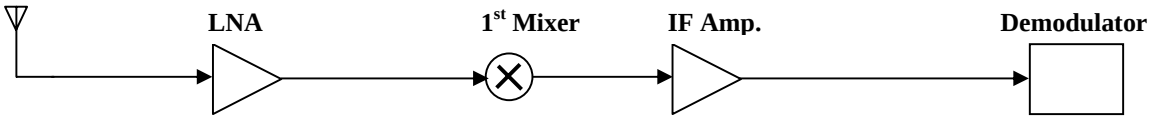
5.4.3 Why "Zero" IF or Direct Conversion

In a direct conversion or so-called "zero IF" communication system, any conceivable DC offset current might screw up the de-modulation of the received signals so that it results the communication system out of work. This is why the differential configuration is always adapted in a direct conversion communication system because its DC offset is theoretically zero. However, recalling all the disadvantages listed in Table 5.3, it would certainly doubt whether it is worthy to build differential circuits for the "zero" IF communication system or not, unless the "zero" IF system brings about huge benefit to business.

Yes, a "zero" IF or direct conversion communication system is much cheaper than a double or triple conversion communication system. A dual conversion system was usually used in the early stages of portable radio or cellular phone development; now a direct conversion system is preferred. Figure 5.35 shows their block diagrams respectively.



(a) Dual conversion receiver



(b) Direct conversion receiver

Figure 5.35 Block diagram of dual conversion and direct conversion receiver

It can be seen that all the filters in the dual conversion system cease to exist in the direct conversion system. It results in a huge benefit in two aspects:

1) Most filters are built up by capacitors and inductors. This is no technological problem in the constructed discrete module; however, it is a big problem in integrated circuits nowadays because the inductors have a very low Q values and occupy large areas in the IC die.
2) The price of a dual conversion system is significantly higher than that of a direct conversion system. For example, a dual-pole IF crystal filter in the dual conversion system costs about $5, which is almost equal to the total cost of a set of IC's for a direct conversion communication system. After taking into account all the other filters, the cost of a direct conversion system could be calculated to be less than 50% percent the cost of a dual system. And the cost of the product, of course, is the first priority of most people.

5.4.4 DC Offset Cancellation

As shown above, in an ideal differential pair, zero DC offset is a very exciting feature. In the actual engineering design, however, an ideal or perfect differential pair does not exist. A symmetrical layout is never perfect because of the slight mismatching tolerance of the parts and processing. This asymmetry results in the DC-offset problem.

On the other hand, DC offset is induced not only due to the non-linearity of devices which we discussed above, but also due to the interference outside the differencial pair which we did not discussed yet. For example, in a differential mixer working for a direct conversion system, the power of LO injection might leak to the LNA output and return back to its RF port so that it causes another type of DC offset at the mixer output.

Consequently, the cancellation of the DC offset is absolutely necessary though the configuration of differential pair is adapted in the circuit design for a direct conversion system. The research and development of this project has been moving forward and many excellent works have achieved lately.

5.4.4.1 "Chopping" Mixer

In a receiver, DC offset can be cancelled or significantly reduced at the mixer block. One of the mixers developed for such a purpose is called the "chopping" mixer, or a dynamically matched differential mixer. By means of "chopping" the input and output signal of the mixer, the even orders of non-linearity of the mixer or the DC offset can be significantly reduced.

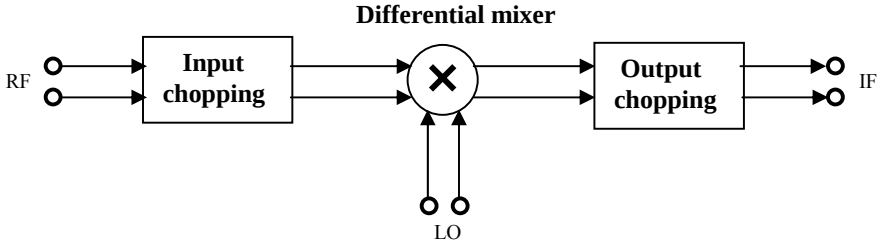
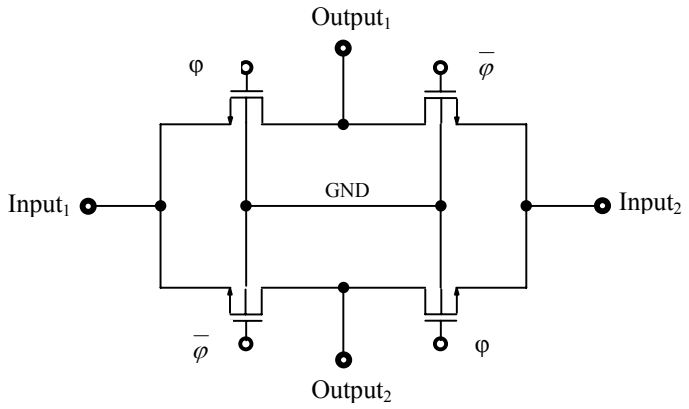


Figure 5.36 Block diagram of a "chopping" mixer

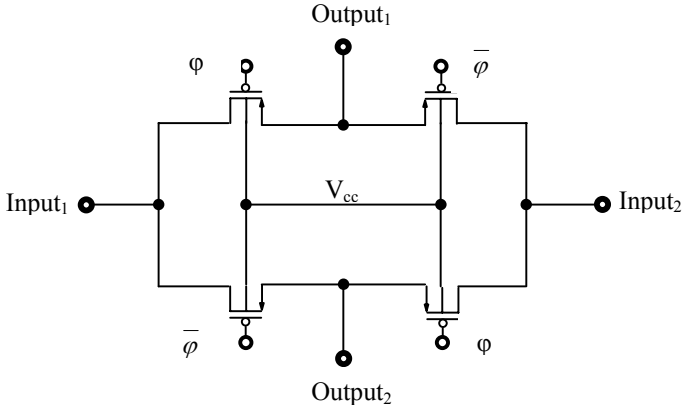
A chopping mixer consists of three portions: an input chopping block, a regular differential mixer, and an output chopping block. It is, in fact, a differential mixer, say, a Gilbert cell, connected to input and output chopping blocks as shown in Figure 5.36.

Figure 5.37 depicts the input and output chopping blocks. Both of them are simply cross-toggled switches. The input chopping block consists of 4 n-channel MOSFET's and its

substrate is grounded while the output chopping block consists of 4 p-channel MOSFET's and its substrate is connected to the DC power supply V_{cc}. Two differential inputs are alternatively connected to two differential outputs through two gate controls φ and $\overline{\varphi}$, which are fed with pulse signals of an appropriate chopping frequency f_{CH} and of the voltage amplitude, 0 and V_{cc}. The phase of φ and $\overline{\varphi}$ is kept at a 180^o difference, that is, φ is at 0 when $\overline{\varphi}$ is at V_{cc}, and vice versa. The chopping frequency is experimentally determined by simulation as well as bench-work. The criterion for determination is to maximize IP2 or to minimize the DC offset for the entire chopping mixer. The IP2 is good for measuring DC offset because it is concerned with the dominating terms among the even order nonlinearities which bring about DC offset. Like the differential mixer itself, the layout of the input and output chopping block must be kept at a good symmetrical configuration as much as possible.



(a) Input chopping block



(b) Output chopping block

Figure 5.37 Input and output chopping blocks of a "chopping" mixer

Figure 5.38 is the schematic of the entire "chopping" mixer. The Gilbert cell is built up with bipolar devices while the input and output chopping blocks are built up with MOSFETs. In the RFIC design it is the advantage of the BiCMOS processing. The input RF signals are fed to the emitters of the lower bipolar pair and are therefore called the CB configuration of the RF input. The advantage of the CB configuration is that the RF input has a relatively wider bandwidth but lower conversion gain than other configurations, such as CE or CC. One of the possible ways to connect two RF input ports is to use an RF transformer. The RF inputs shown in Figure 5.38 are connected to the second winding of the transformer with its center tapped to the ground while the primary winding of the transformer could be either differential or single-ended RF input.

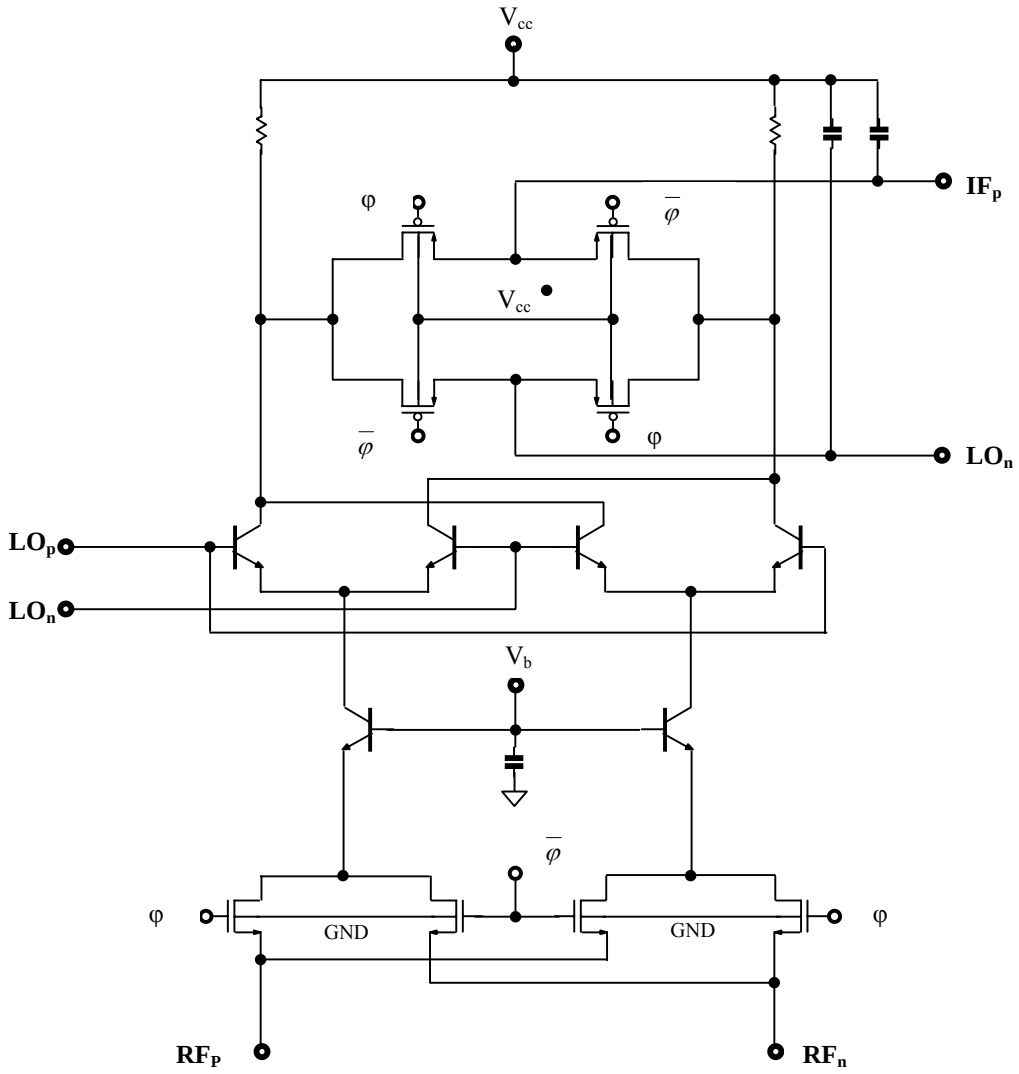


Figure 5.38 Topology of a "chopping" mixer

Now, it is desirable to understand why the "chopping" technology can cancel or significantly reduce DC offset.

DC offset is contributed by the even orders of the non-linearity of a device or block. The second order of non-linearity is, of course, by far the most effective contributor. It is therefore reasonable to take IP_2 as a measure of the DC offset for a device, a block, or a system.

The even orders of the non-linearity of a device or block are the "land" of DC offset production while the interferers of the signal are the "seeds" of the DC offset production. There are two kinds of interferers that can bring about the DC offset:

1) Two interferers with their frequency difference, Δf.
As shown in Figure 5.39, assuming that the two interferers are

$$S_{i1} = S_{io} \cos \omega_1 t \ , \tag{5.179}$$

$$S_{i2} = S_{io} \cos \omega_2 t \ , \tag{5.180}$$

and

$$\Delta f = \frac{\omega_1 - \omega_2}{2\pi} \ , \tag{5.181}$$

where S_{i1} and S_{i2} = Interferer 1 and 2 respectively,
ω_1 and ω_1 = Angular frequency of interferer 1 and 2 respectively,
S_{io} = Amplitude of interferer 1 or 2.

It should be noted that the frequencies of two interferers could be in a very high frequency range and their frequency difference could be in any value.

If these two interferes enter into the mixer, then the 2nd order of the non-linearity of the mixer produces the DC offset and the un-desired 2nd order spurious product, S_{u2}, is produced:

$$S_{u2} = a_2 \left(S_{i1} + S_{i2} \right) = a_2 \left(S_{io} \cos \omega_1 t + S_{io} \cos \omega_2 t \right)^2$$

$$S_{u2} = a_2 S_{io}{}^2 \left[1 + \frac{1}{2} \left(\cos 2\omega_1 t + \cos 2\omega_2 t \right) + \cos 2\pi \Delta f t + \cos\left(\omega_1 + \omega_2 \right) t \right], \tag{5.182}$$

where a_2 = Coefficient of 2nd order non-linearity of the device or block.

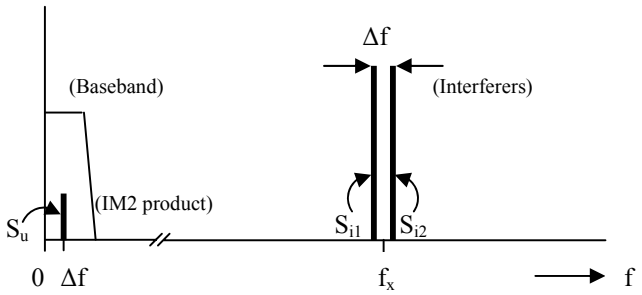


Figure 5.39 Two interferers result an un-desired IP2 product in a mixer.

On the right side of expression (5.182), the 1st term is the DC offset and the 3th term is the 2nd order spurious product. The expression (5.182) implies that the 2nd order spurious product, *cos2πΔft*, is a good measure of the DC offset since they have the same amplitude, $a_2 S_{io}^2$.

2) One interferer with high power

As long as the non-linearity exists, a single interferer with high power could bring about DC offset.

If the interferer is

$$S_i' = S_{io}' \cos \omega t \ , \tag{5.183}$$

Then

$$S_{u2}' = a_2 S_i'^2 = a_2 S_{io}'^2 \left(1 + \frac{1}{2}\cos 2\omega t\right), \tag{5.184}$$

where S_i'= Interferer,
ω = Angular frequency of interferer,
S_{io}' = Amplitude of interferer.

The 1st term in expression (5.184) is the DC offset. It should be noted that the frequency of interferer could be of any value.

In both cases above, more DC offset could be produced if the amplitude of the interferers is not constant. For instance, if the amplitude of the interferers is a time-dependent, that is, if

$$S_{io} = S_o\left[1 + m(t)\right] , \tag{5.185}$$

or

$$S_{io}' = S_o'\left[1 + m'(t)\right] , \tag{5.186}$$

then, expressions (5.182) and (5.184) become

$$S_{u2} = a_2 S_o{}^2 [1+m(t)]^2 \left[1+\frac{1}{2}(\cos 2\omega_1 t + \cos 2\omega_2 t) + \cos 2\pi\Delta f t + \cos(\omega_1+\omega_2)t\right],$$

$$S_{u2} = a_2 S_o{}^2 [1+2m(t)+m^2(t)]\left[1+\frac{1}{2}(\cos 2\omega_1 t + \cos 2\omega_2 t) + \cos 2\pi\Delta f t + \cos(\omega_1+\omega_2)t\right] \quad (5.187)$$

$$S_{u2}' = a_2 S_i'^2 = a_2 S_o'^2 [1+2m'(t)+m'^2(t)]\left(1+\frac{1}{2}\cos 2\omega t\right), \quad (5.188)$$

respectively.

It can be seen that a time-independent DC offset,

$$a_2 S_{io}{}^2 [m(t)+ 2m^2(t)], \quad \text{or} \quad a_2 S_{io}{}^2 [m'(t)+ 2m'^2(t)],$$

is included, which is troublesome because it is very hard to control.

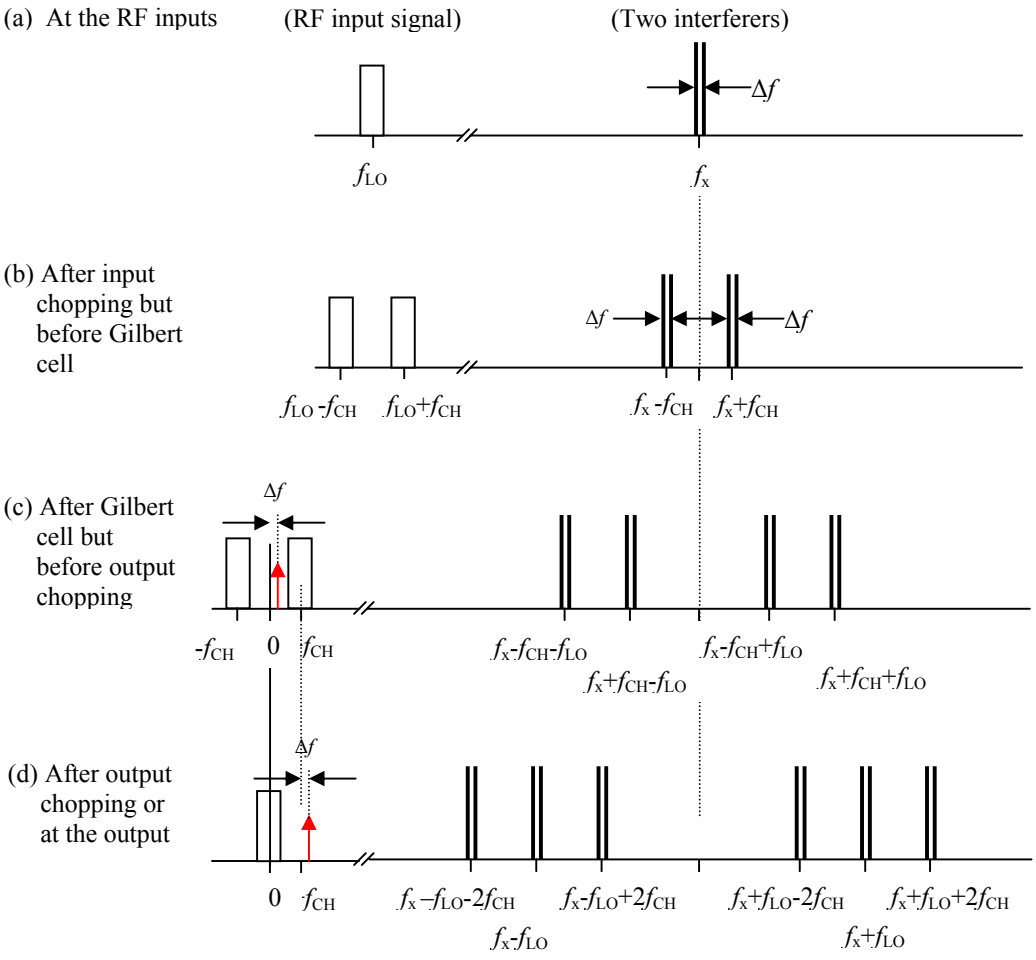


Figure 5.40 Principle of DC offset reduction in the "Chopping" mixer

Now let's return to the case of the chopping mixer.

It is assumed that the chopping mixer, as shown in Figure 5.38, is working for the "zero" conversion or direct conversion of the input RF signal. Therefore, at the RF differential inputs, there is an RF input signal with a LO frequency. In addition to the input RF signal, let's assume that there are two interferers with an average frequency of f_x and a frequency difference of Δf. Figure 5.40 (a) depicts them together.

The input chopping block functions like a mixer. The chopping signal operates like a local injection of a mixer and makes the frequency spectrum at the RF input "split" into two components with a positive and a negative chopping frequency shift. Figure 5.40 (b) shows the resulting frequency spectrum after input chopping.

Based on the same principle, Figure 5.40 (c) shows the spectrum after the Gilbert cell but before the output chopping, in which the spectrum as shown in Figure 5.40 (b) is "split" into two components with a positive and a negative LO frequency shift. It should be noted that the spectrum around the higher frequencies, $2f_{LO}$-f_{CH} and $2f_{LO}$+f_{CH}, is ignored since we are only interested in the spectrums that could be converted to a low frequency. Special attention must be paid to the 2^{nd} order product of two interferers, which is represented by a vertical arrow in Figure 4.40 (c). Its frequency is Δf and would be inside the baseband when Δf is less than the frequency bandwidth of the baseband.

However, at the output of the "chopping" mixer or after the output chopping block, the desired RF signal returns to the baseband. And, as long as the f_{CH} is higher than the maximum of the baseband frequencies, the 2^{nd}-order product of two interferers is "pumped" up outside the baseband so that the goal of the chopping scheme is reached. This is shown in the Figure 5.40 (d), in which the spectrum with higher frequencies is ignored again.

Undoubtedly, the "chopping" technology is an effective means to cancel or significantly reduce the DC offset in a "zero" IF or direct conversion communication system. However, the chopping signal might turn into a new source of disturbance itself since it spreads over the mixer block. Special attention should be paid to this.

5.4.4.2 DC Offset Calibration

Another way to accomplish DC offset cancellation is by means of DC offset calibration. As long as the DC offset is kept at a minimum during the system operating period, the DC offset would affect on the de-modulation of the desired signal from the carrier at a minimum level though it is a "zero" IF or a direct conversion system.

Figure 4.41 shows a simplified block diagram of DC offset calibration, which is developed by Intersil in their baseband IC, HFA3783, for WLAN products. To keep the DC offset constant, the combination of all DC offsets produced by the mixer, LPF (Low Pass Filter), and the buffer must be calibrated at any time during the receiving, transmitting, or power down mode.

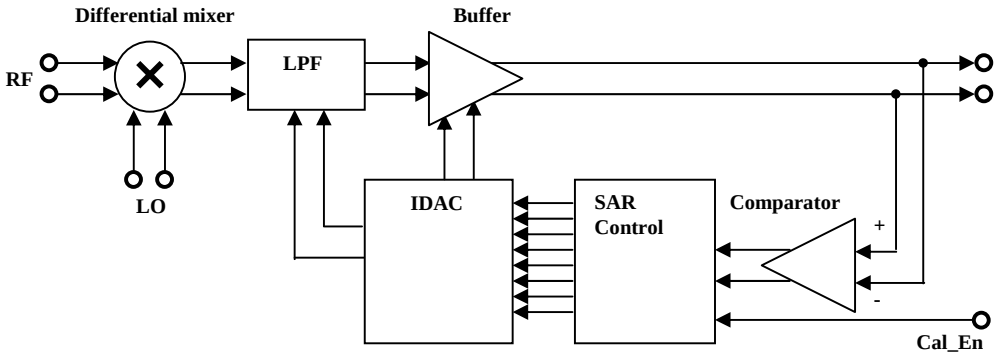


Figure 5.41 Simplified block diagram of DC offset calibration in Intersil's baseband IC, HFA3783.

The comparator senses the DC offset from the buffer output and its output is fed to a decision circuit which changes the condition of a SAR (Successive Approximation Register) state control. The 8 digital outputs of SAR control the IDAC (current output of Digital to Analog Conversion) which adjusts and brings the combination of all DC offset to a minimum. The DC offset calibration is initialized by the Cal_En (Calibration Enable) input either automatically or manually.

As we can see, the technology of the DC offset calibration is somehow complicated and might bring about new types of spurious products since there are many digital signals involved.

5.4.4.3 Hardware Schemes

In the technologies of the DC offset cancellation, "chopping" is an active means because it eliminates the DC offset before it turns up. The DC offset calibration is a somehow passive means because it eliminates the DC offset after it turns up. Both of the technologies are somehow complicated and might bring about new problems in the actual engineering design.

A simple idea to cancel DC offset is to insert a DC-blocking capacitor before the de-modulation block. This, however, is impractical because for it to work, the value of the capacitor needs to be ridiculously high: infinitive. Another idea is to insert a high pass filter and a low frequency amplifier before the de-modulation block as shown in Figure 5.42.

The value of the capacitance of the capacitor in the HPF needs not be an unreasonably high value when the HPF is built by a combination of capacitors and resistors. The low-frequency amplifier is needed because such an HPF causes a significant loss of the desired signal. The question is what the 3 dB corner frequency of the HPF should be. From the viewpoint of the fidelity of desired signal, the 3 dB corner frequency must be as low as possible. The lower the 3 dB corner frequency, the higher the capacitor value in the HPF must be. A trade-off between the 3 dB corner frequency and the fidelity of the desired signal must be made.

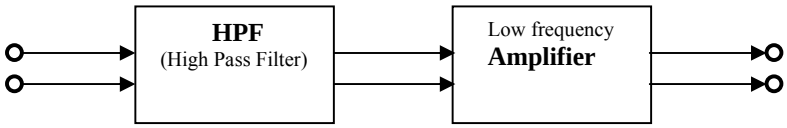


Figure 5.42 DC offset cancelled by the combination of a low-frequency amplifier and an HPF.

References

[1] George D. Vendelin; Anthony M. Pavio; Ulrich L. Rohde; “Microwave Circuit Design Using Linear and Nonlinear Techniques” (Book), John Wiley & Sons, Inc., 1990.

[2] Richard C. Li; “Double Line RF Balun Transformer,” U. S. Patent 5,477,204, Motorola Inc., 1993.

[3] Paul R. Gray, Paul J. Hurst, Stephen H. Lewis, and Robert G. Meyer, “Analysis and Design of Analog Integrated Circuits,” (Book), Fourth Edition, John Wiley & Sons, Inc., 2001.

[4] Shoval, A.; Johns, D.A.; Snelgrove, W.M.; “DC offset performance of four LMS adaptive algorithms,” 1994 IEEE International Symposium on Circuits and Systems, 1994. ISCAS '94, Vol. 2 , pp. 409 – 412, May 30 - June 2, 1994

[5] Shoval, A.; Johns, D.A.; Snelgrove, W.M.; “Comparison of DC offset effects in four LMS adaptive algorithms,” IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, [see also IEEE Transactions on Circuits and Systems II: Express Briefs,] Vol. 42 , Issue 3 , pp. 176-185, March 1995.

[6] Yoshida, H.; Tsurumi, H.; Suzuki, Y.; “DC offset canceller in a direct conversion receiver for QPSK signal reception,” The Ninth IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, 1998., Vol. 3 , pp. 1314 – 1318, Sept. 8-11, 1998.

[7] Edwin E. Bautista, Babak Bastini, and Joseph Heck, “A High IIP2 Down-conversion Mixer Using Dynamic Matching”, IEEE Journal of Solid-state Circuits, Vol.35, NO.12, pp 1934-1941, December, 2000.

[8] Matinpour, B.; Chakraborty, S.; Laskar, J.; “Novel DC-offset cancellation techniques for even-harmonic direct conversion receivers,” IEEE Transactions on Microwave Theory and Techniques, Vol. 48, Issue12 , pp.2554 – 2559, Dec. 2000.

[9] Matinpour, B.; Chakraborty, S.; Hamai, M.; Chun, C.; Laskar, J.; “A novel dc-offset cancellation technique for even-harmonic direct conversion receivers,” Microwave Symposium Digest., 2000 IEEE MTT-S International , Vol. 2 , pp. 631 – 634, June 11-16, 2000.

[10] Po-Chiun Huang, Yi-Huei Chen, and Chorng-Kuang Wang, “A 2-V CMOS 455-kHz FM/FSK Demodulator Using Feedforward Offset Cancellation Limiting Amplifier”, IEEE Journal of Solid-state Circuits, Vol.36, NO.1, pp 135-138, January, 2001.

[11] Faulkner, M.; “IM2 removal in direct conversion receivers,” IEEE VTS 53rd Vehicular Technology Conference, 2001. VTC 2001 Spring., Vol. 3 , pp. 1897 – 1901, May 6-9, 2001.

[12] Faulkner, M., “DC offset and IM2 removal in direct conversion receivers; ” Communications, IEE Proceedings, Vol. 149, Issue 3, pp. 179 – 184, June 2002.

[13] Popov, O.V.; “Dynamic DC offset impact on the 802.11a receiver performance,” Proceedings of ICCSC '02. 1st IEEE International Conference on Circuits and Systems for Communications, 2002. , pp. 250-253, June 26-28, 2002.

[14] Xinping Huang; Caron, M.; “Gain/phase imbalance and DC offset compensation in quadrature modulators,” 2002 IEEE International Symposium on Circuits and Systems, ISCAS 2002., Vol. 4, pp. IV-811 - IV-814, May 26-29, 2002.

.

[15] Shang, S.; Mirabbasi, S.; Saleh, R.; “A technique for DC-offset removal and carrier phase error compensation in integrated wireless receivers,” Proceedings of the 2003 International Symposium on Circuits and Systems, 2003. ISCAS '03. Proceedings of the 2003 International Symposium on Circuits and Systems, Vol.1 , pp. I-173 - I-176, May 25-28 , 2003.

Index

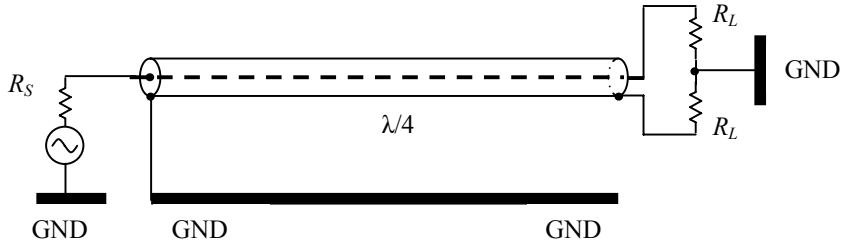
Contents

Chapter 6 Balun

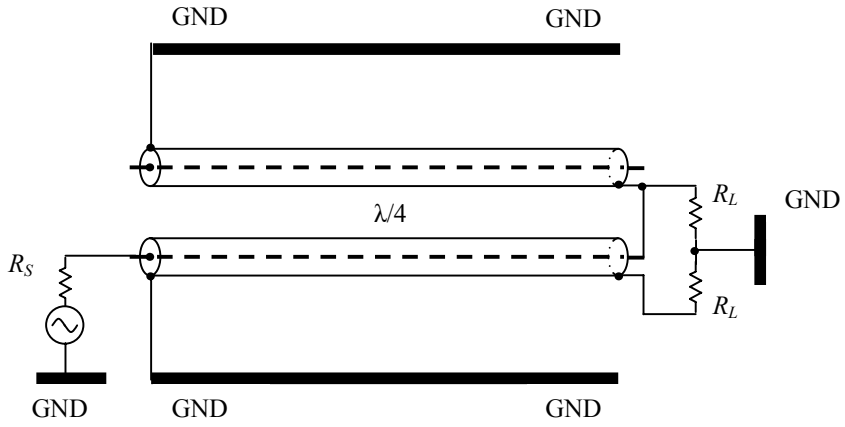
A balun performs transformation between single-ended stage and differential pair. A variety of baluns have been developed in the past decades. Only some of them are introduced here.

6.1 Coaxial Cable Balun

At an early stage in the development of the microwave system, a cable line was applied as a simple balun. Figure 6.1 shows how the coaxial cable bulun is connected with the source and the load.



(a) Single coaxial line balun



(b) Double coaxial line balun

Figure 6.1 A balun made by coaxial line

The electric length of the coaxial cable is a quarter wavelength, $\lambda/4$, so that the phase of two output balance ports is 90^o and -90^o respectively. Two balanced outputs in Figure 6.1 (a) might be somehow unbalanced due to the imperfect balance between the central wire and the outer cylindrical conductive surface. Therefore, another coaxial cable line is added beside the main coaxial line in parallel and its outer cylindrical conducting surface is connected to the central wire of the main coaxial cable as shown in Figure 6.1(b), so that the balance is improved at the two output balanced ports.

The advantages of this type of balun are:
1) It is simplest and cheapest,
2) The insertion loss is usually only several tenths dB, which is negligible.

The disadvantages of this type of balun are:
1) Its frequency response is narrow because its length is confined to a quarter wavelength around the operating frequency.
2) The length of a coaxial cable corresponding to the quarter wavelength depends on the type, material, and the diameter of the cable. For example, in an RF range, say, 500 MHz, it is abut several inches. It implies that the coaxial cable balun can be applied only to test boards with a somehow large size.

6.2 Ring Micro Strip Line Balun

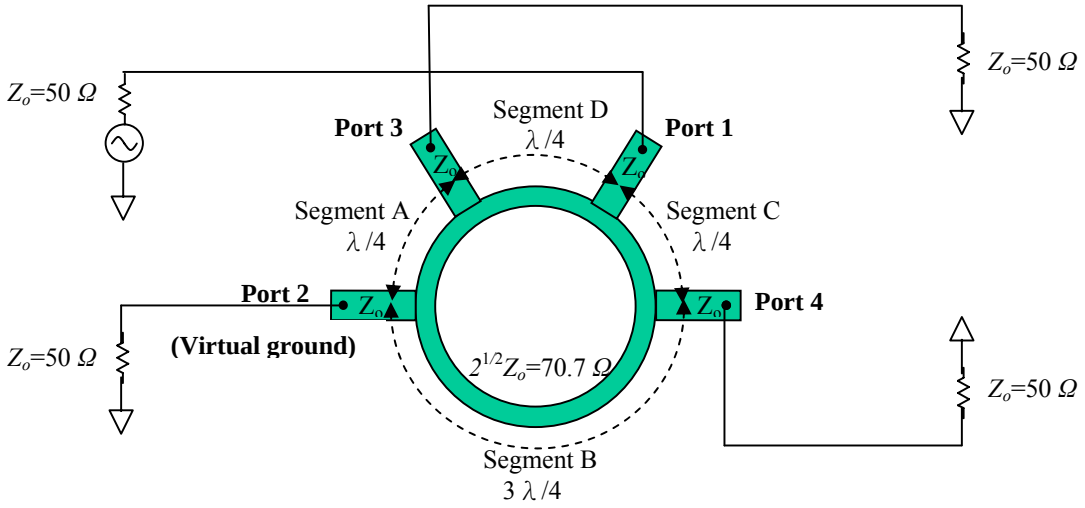


Figure 6.2 Layout of "Ring" or "Rat-race" balun

Ring or "rat-race" balun is another type of micro strip line balun.

As shown in Figure 6.2, the ring's circumference must be equal to 1.5 wavelengths of the electric length, and its width must be chosen so that its characterized impedance is equal to $2^{1/2}Z_o$. The characterized impedance of the 4 tapes must be equal to the source or load impedance Z_o. The arc of the ring is one-quarter wavelength from port 2 to port 3, port 3 to port 1, or port 1 to port 4. The arc of the ring is three quarter wavelengths from port 4 to port 2.

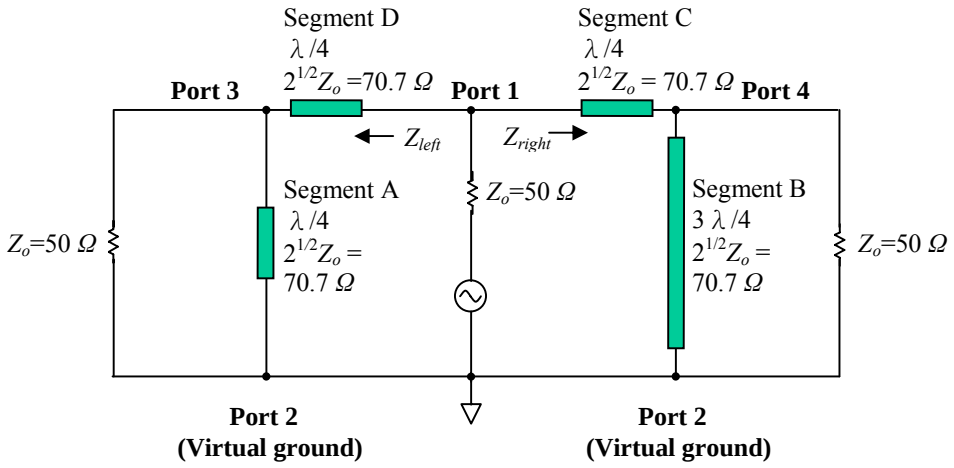
If port 1 is chosen as single-ended input port, then port 2 is a virtual grounding because it is distanced from port 1 by one-half wavelength. The phase in ports 3 and 4 then are 90^o and -90^o respectively since they are a quarter wavelength forward and backward distant from port 1 correspondingly.

In order to explain the impedance matching, the equivalent circuit of the "Ring" or "Rat-race" balun for all the segments is plotted in Figure 6.3 (a). The impedance of segment B looking from port 4 is infinitive because its length is 3 quarters of the wavelength from the virtual grounding port 2.

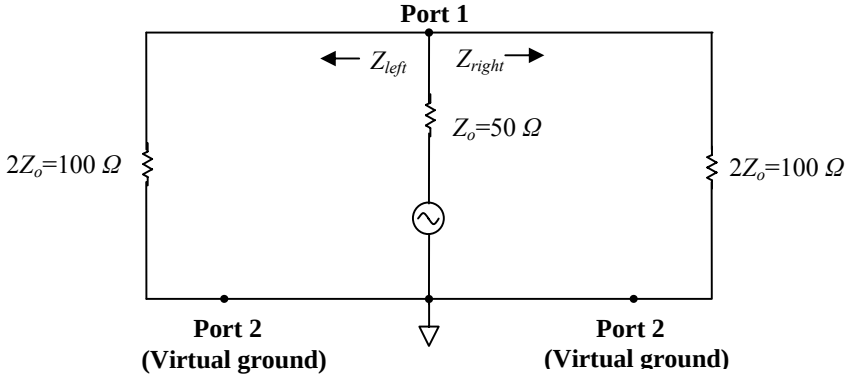
Similarly, the impedance of segment A is also infinitive because its length looking from port 3 is one quarter of the wavelength from the virtual grounding port 2. Consequently the segments B and A can be neglected in the examination of the impedance matching as shown in Figure 6.3 (b). Now, let's look the impedance at point 4 on the right side or left side. Based on the transmission line theory, for a segment with quarter wavelength, the relation of the impedances between two terminals and the characteristic impedance is

$$Z_{right} = Z_{left} = \frac{\left(2^{0.5} Z_o\right)^2}{Z_o} = 2Z_o \quad , \qquad (6.1)$$

where Z_{right} or Z_{left} = the impedance at one end of the segment with 1 quarter wavelength,
Z_o = the impedance at another end of the segment with 1 quarter wavelength,
$2^{1/2} Z_o$= the characteristic impedance of the micro strip segment.



(a) 1st equivalent circuit with all segments



(b) 2nd equivalent circuit to display impedance matching

Figure 6.3 Equivalent circuit of "Ring" or "Rat-race" balun.

Then the impedance looking outward from port 1 is

$$Z_{port1,out} = Z_{right} // Z_{left} \ . \tag{6.2}$$

From expression (6.1),

$$Z_{port1,out} = Z_o \ . \tag{6.3}$$

It is therefore concluded that the looking outward impedance at port 1 is matched to the source impedance at port 1, Z_o, if Z_o is a pure resistor.

The “Ring” or “Rat-race” balun is a narrowband. Its relative bandwidth is usually less than 15%.

6.3 Transformer Balun

There are many ways to build a transformer balun, such as a transformer constructed on a binocular ferrite core with trifilar windings, a transformer constructed on a ceramic block with high permeability or permittivity, the spiral inductors with tight coupling on an integrated circuit, and so on. We are going to introduce the transformer that is constructed on an RFIC chip by spiral inductors.

Table 6.1 illustrates the difference between a transformer built on a ferrite core with copper windings and on an RFIC chip with spiral inductors. The merit of a ferrite transformer is the high Q value of the inductor. Despite other merits and demerits, its fatal weakness is that the maximum operating frequency is less than about 1 GHz. At present, the cost is the first priority to be considered in the circuit design. Consequently there seems to be no choice but to build the transformer on the RFIC chip although its Q value is pretty low.

Table 6.1 Comparison between a ferrite transformer and a spiral IC transformer

	Ferrite transformer	Spiral IC transformer
Max. frequency	<1 GHz	~ 10 GHz
Cost	High	Low
Space/Area	Large	Small
Q value	Acceptable (>10)	Too low (<10)
Symmetry	Good	Good
Bandwidth	OK	OK

There are approximately 3 basic types of transformers on an RFIC chip with a spiral configuration.

1) Tapped transformer

A tapped transformer is built on the top metal layer as shown in Figure 6.4. The primary and secondary windings are formed by the inner spiral and the outer spiral. The spacing between two adjacent wires, *S,* is usually 1 micrometer. This is not only for the IC die area saving but also for the Q value enhancing. The width of the wire, *W*, is typically 10 to 50 micrometers. Too narrow of a wire would result in a low Q value and too wide of a wire would be not good to the IC die area saving. The inner diameter, *D*, is an important parameter in the IC transformer design. From the view point of high Q value, *D* should be as high as possible. However, a higher *D* might make the IC die area high to an un-affordable size.

For this type transformer, one of advantages is the low capacitance between the primary and secondary windings. However, the main disadvantage is the weak coupling. The coupling coefficient, *k*, is only in the range from 0.3 to 0.5. Another disadvantage is that the asymmetrical configuration makes impossible to build it for the differential pair.

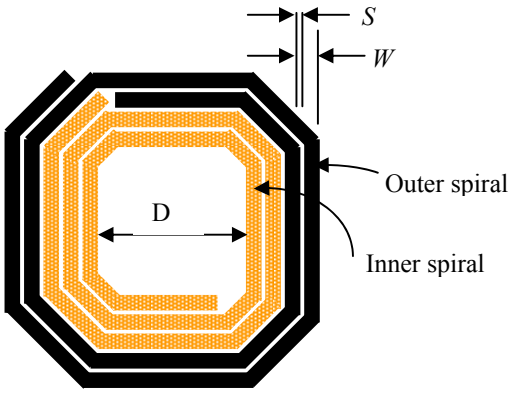


Figure 6.4 A tapped spiral transformer

2) Interleaved transformer

An interleaved transformer is built on the top metal layer also. The primary and secondary windings are interleaved from each other as shown in Figure 6.5.

For this type transformer, the main advantage is that the good symmetry makes it a good candidate to use in differential circuits. However, its capacitance between the primary and secondary windings is greater and the coupling coefficient is higher than that of tapped transformer. The k is usually in the range from 0.7 to 0.8.

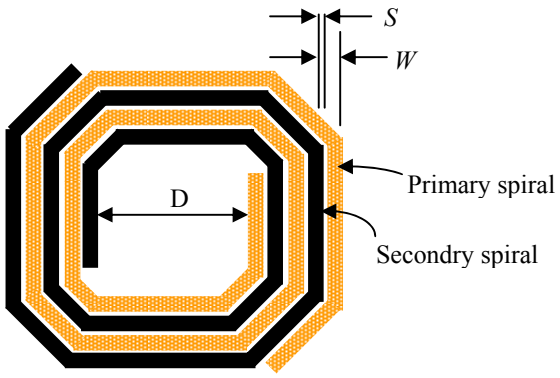


Figure 6.5 An interleaved spiral transformer

3) Stacked transformer

In order to build a stacked transformer in a IC chip as shown in Figure 6.6, the processing must have two metallic layers at least. As shown in Figure 6.6, the primary and secondary

windings are on the top metal layer and the next layer beneath the top layer respectively. Actually these two windings are overlapped as exactly as possible. In the drawing of Figure 6.9, we intentionally shift these two windings a little bit so that readers can still see the winding in the bottom layer somehow.

For this type of transformer, the main advantage is that the strongest coupling can be obtained. The coupling coefficient *k* is about 0.9. Also, it is the best way for IC die size saving. However, the capacitance between the primary and secondary windings is higher than that of a tapped or interleaved transformer.

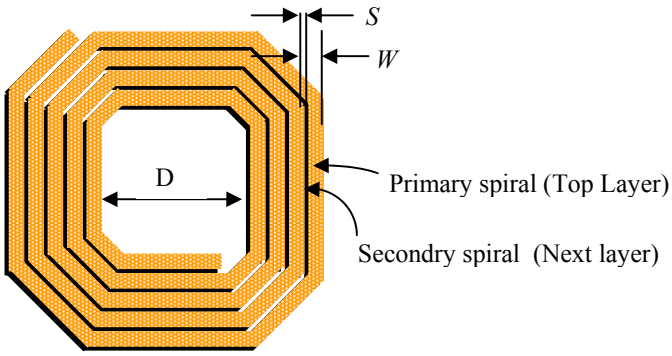


Figure 6.6 A stacked spiral transformer

Table 6.2 makes a comparison of the main parameter of these three types of IC transformers.

Table 6.2 Comparison between IC transformers

Transformer type	IC die area	*k*	Capacitance	*SRF*
Tapped	High	Low	Low	High
Interleaved	High	Middle	Middle	High
Stacked	Low	High	High	Low

Notes : *k* = Coupling coefficient;
SRF = Self-Resonant Frequency.

6.4 Transformer Balun Composed by Two Stacked 2x2 Transformers

Usually an IC foundry provides models of the spiral transformers. The designer can build a balun by two 2x2 transformers stacked together, where "2x2" means 2 ends of input and 2 ends of output. Figure 6.7 to 6.10 shows 4 ways that one can implement the transformer balun based on what you need.

Figure 6.7 shows a transformer balun constructed by two stacked transformers with their turn ratio, 1:*n*. It is a balun from single-ended to differential. At the single-ended portion, two primary windings are connected in series and at the connected point, the 1 GΩ resistor is meaningless in the practical peformance but just for the DC grounding purpose in the simulation. At the differential portion, two secondary windings are connected in series and its connected point is the central grounded point of the balun. This way of connection is denoted with (*ss-ds*), which means that "**S**ingle-ended, primary windings stacked in **S**eries, goes to **D**ifferential, secondary windings stacked in **S**eries."

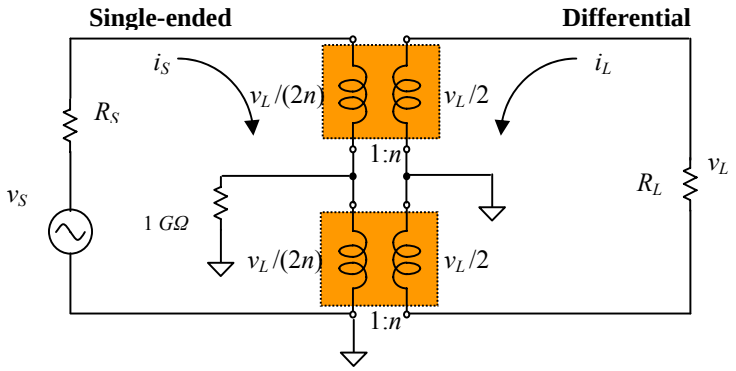


Figure 6.7 A transformer balun by two stacked transformers (*ss_ds*)

From Figure 6.7, we have

$$v_S = \frac{2v_L}{2n} = \frac{v_L}{n} = \frac{i_L R_L}{n} = \frac{i_S}{n}\frac{R_L}{n} = \frac{i_S R_L}{n^2} \quad , \tag{6.4}$$

then,

$$R_S = \frac{v_S}{i_S} = \frac{R_L}{n^2} \quad . \tag{6.5}$$

Based on the values of R_S and R_L, the ratio n can be calculated from expression (6.5).

Figure 6.8 shows the transformer balun constructed by two staked transformers with their turn ratio, 1:n. It is a balun from single-ended to differential. At the single-ended portion, two primary windings are connected in parallel and its connected point is grounded. At the differential portion, two secondary windings are connected in series and its connected

point is the central grounded point of the balun. This way of connection is denoted with (*sp-ds*), which means that "**S**ingle-ended, primary windings stacked in **P**arallel, goes to **D**ifferential, secondary windings stacked in **S**eries."

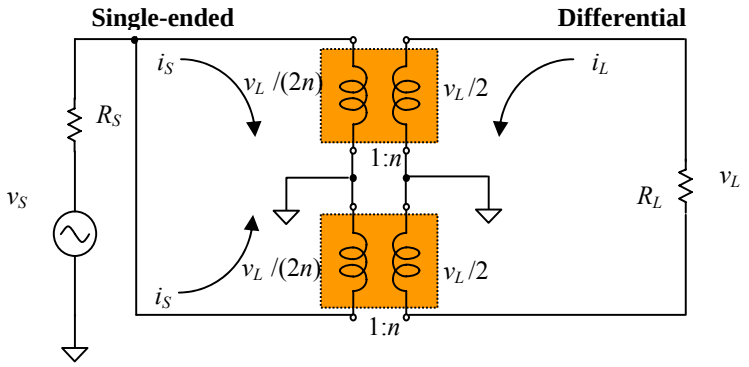


Figure 6.8 A transformer balun by two stacked transformers (*sp_ds*)

From Figure 6.8, we have

$$v_S = \frac{v_L}{2n} = \frac{i_L R_L}{2n} = \frac{i_S R_L}{2n^2} \ , \tag{6.6}$$

then,

$$R_S = \frac{v_S}{2i_S} = \frac{R_L}{4n^2} \ . \tag{6.7}$$

Based on the values of R_S and R_L, the ratio n can be calculated from expression (6.7).

Figure 6.9 shows the transformer balun constructed by two staked transformers with their turn ratio 1:n. It is a balun from differential to single-ended. At the differential portion, two primary windings are connected in series and its connected point is grounded. At the single-ended portion, two secondary windings are connected in series and 1 GΩ resistor to the connected point is meaningless in the practical peformance but just for the DC grounding purpose in the simulation.. This way of connection is denoted with (*ds-ss*), which means that "**D**ifferential, primary windings stacked in **S**eries, goes to **S**ingle-ended, secondary windings stacked in **S**eries."

From Figure 6.9, we have

$$v_S = \frac{2v_L}{2n} = \frac{v_L}{n} = \frac{i_L R_L}{n} = \frac{i_S}{n}\frac{R_L}{n} = \frac{i_S R_L}{n^2} \ , \tag{6.8}$$

then

$$R_S = \frac{v_S}{i_S} = \frac{R_L}{n^2} \ . \tag{6.9}$$

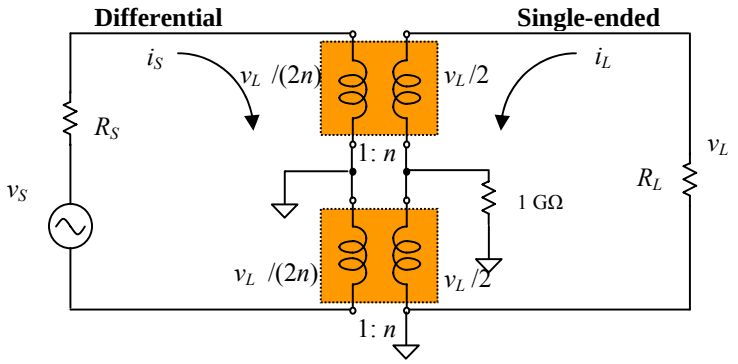


Figure 6.9 A transformer balun by two stacked transformers (*ds_ss*)

Based on the values of R_S and R_L, the ratio n can be calculated from expression (6.9).

Figure 6.10 shows the transformer balun constructed by two staked transformers with their turn ratio 1:n. It is a balun from differential to single-ended. At the differential portion, two primary windings are connected in series and its connected point is grounded. At the single-ended portion, two secondary windings are connected in parallel. This way of connection is denoted with (*ds-sp*), which means that "**D**ifferential, primary windings stacked in **S**eries, goes to **S**ingle-ended, secondary windings stacked in **P**arallel."

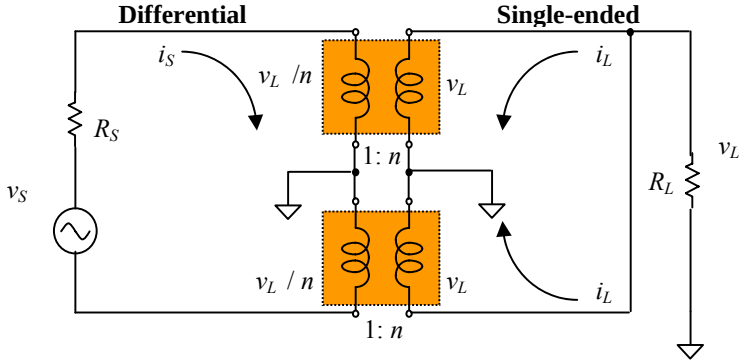


Figure 6.10 A transformer balun by two stacked transformers (*ds_sp*)

From Figure 6.10, we have

$$v_S = \frac{2v_L}{n} = 2\frac{2i_L R_L}{n} = 2\frac{2\frac{i_S}{n}R_L}{n} = 4\frac{i_S R_L}{n^2} \quad , \qquad (6.10)$$

then,

$$R_S = \frac{v_S}{i_S} = 4\frac{R_L}{n^2} \quad . \tag{6.11}$$

Based on the values of R_S and R_L, the ratio n can be calculated from expression (6.11).

6.5 LC Balun

An LC balun can be made of discrete parts, chip parts, MLC (Multi-Layer-Ceramic), or IC chips. It is an easy and fast way to build an LC balun by chip parts on a PCB for circuit implementation or for testing.

The simplest LC balun consists of only two identical inductors and two identical capacitors. These 4 parts form a bridge structure as shown in Figure 6.11. It is shown that the core of the LC balun can serve as a splitter or a combiner. The following analysis will be aimed at the splitter since it is similar at the combiner.

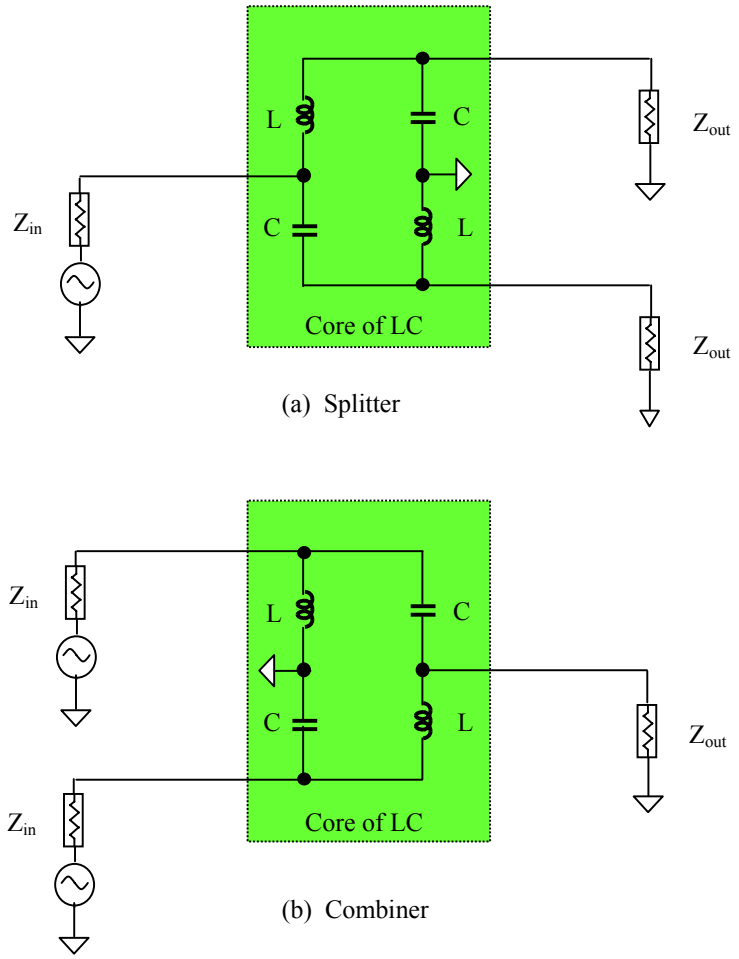


Figure 6.11 LC balun functions as a splitter or a combiner

There is a single port with its impedance Z_{in} and two balanced ports with their impedance Z_{out}. They are

$$Z_{in} = R_{in} + jX_{in} \ , \qquad (6.12)$$

$$Z_{out} = R_{out} + jX_{out} \ , \tag{6.13}$$

where Z_{in} = Single port impedance,
R_{in} = Real part of input port impedance,
X_{in} = Imaginary part of input port impedance,
Z_{out} = Balanced port impedance,
R_{out} = Real part of output port impedance,
X_{out} = Imaginary part of output port impedance.

For the analysis, let's re-draw Figure 6.11(a) to Figure 6.12.

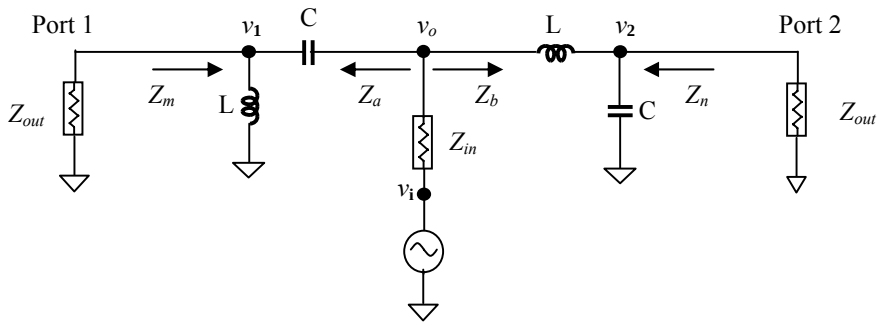


Figure 6.12 Re-drawn the splitter of Figure 6.11

In Figure 6.12, the impedances looking into right hand side and left hand side from the single or input port are

$$Z_a = Z_c + Z_L \,//\, Z_{out} \ , \tag{6.14}$$

$$Z_b = Z_L + Z_c \,//\, Z_{out} \ , \tag{6.15}$$

where

$$Z_c = \frac{1}{j\omega C} \ , \tag{6.16}$$

$$Z_L = j\omega L \ , \tag{6.17}$$

and ω = Operating angular frequency.

Then,

$$Z_a = \frac{\left(1-\omega^2 LC\right)Z_{out} + j\omega L}{-\omega^2 LC + j\omega C Z_{out}} \ , \tag{6.18}$$

$$Z_b = \frac{\left(1-\omega^2 LC\right)Z_{out} + j\omega L}{1 + j\omega C Z_{out}} \ , \tag{6.19}$$

These two impedances are connected together in parallel from the single or input port to the ground. The resulting impedance must be matched and is equal to the conjugated value of Z_{in}, that is,

$$Z_{in}^{*} = R_{in} - jX_{in} = \frac{Z_a Z_b}{Z_a + Z_b} \ . \tag{6.20}$$

By substituting (6.18) and (6.19) into (6.20), we have

$$Z_{in}^{*} = \frac{\left(1-\omega^2 LC\right)Z_{out} + j\omega L}{\left(1-\omega^2 LC\right) + j2\omega C Z_{out}} \ . \tag{6.21}$$

$$Z_{in}^{*} = \frac{\left(1-\omega^2 LC\right)R_{out} + j\left[\left(1-\omega^2 LC\right)X_{out} + \omega L\right]}{\left(1-\omega^2 LC - 2\omega C X_{out}\right) + j2\omega C Z_{out}} \ . \tag{6.22}$$

From (6.20) and (6.22),

$$R_{in} = \frac{\left(1-\omega^2 LC\right)R_{out}\left(1-\omega^2 LC - 2\omega C X_{out}\right) + 2\omega C R_{out}\left[\left(1-\omega^2 LC\right)X_{out} + \omega L\right]}{\left(1-\omega^2 LC - 2\omega C X_{out}\right)^2 + \left(2\omega C R_{out}\right)^2} \ . \tag{6.23}$$

$$X_{in} = \frac{2\omega C R_{out}\left(1-\omega^2 LC\right)R_{out} - \left(1-\omega^2 LC - 2\omega C X_{out}\right)\left[\left(1-\omega^2 LC\right)X_{out} + \omega L\right]}{\left(1-\omega^2 LC - 2\omega C X_{out}\right)^2 + \left(2\omega C R_{out}\right)^2} \ . \tag{6.24}$$

The expressions (6.22), (6.23), and (6.24), are somehow tedious. However, it can be simplified if

$$\omega^2 LC = 1 \ . \tag{6.25}$$

This condition is expected. The power can be well transmitted from the single port to the balanced ports or vice versa only when the balun core is a resonant-type-core. Under condition(6.25), expressions (6.18), (6.19), (6.22), (6.23), and (6.24) can be simplified as

$$Z_a = \frac{j\omega L}{-1 + j\omega C Z_{out}} \ , \tag{6.26}$$

$$Z_b = \frac{j\omega L}{1 + j\omega C Z_{out}} \ , \tag{6.27}$$

$$Z_{in}^{*} = \frac{L}{2CZ_{out}} , \tag{6.28}$$

$$R_{in} = \frac{L}{2C}\frac{R_{out}}{\left(R_{out}^{2} + X_{out}^{2}\right)} , \tag{6.29}$$

$$X_{in} = \frac{L}{2C}\frac{X_{out}}{\left(R_{out}^{2} + X_{out}^{2}\right)} , \tag{6.30}$$

From expressions (6.25) and (6.28), two conclusive expressions listed below are obtained.

$$\omega L = \sqrt{2Z_{out}Z_{in}^{*}} , \tag{6.31}$$

$$\omega C = \frac{1}{\sqrt{2Z_{out}Z_{in}^{*}}} . \tag{6.32}$$

These two equations are very useful to the design because the values of L and C can be easily calculated from the given impedances Z_{out} and Z_{in}.

Now, let's examine the impedance looking from the balanced ports, Z_m and Z_n.

$$Z_m = \left(Z_b // Z_{in} + Z_c\right) // Z_L , \tag{6.33}$$

$$Z_n = \left(Z_a // Z_{in} + Z_L\right) // Z_c . \tag{6.34}$$

From expressions (6.14), (6.15) and (6.31), (6.32),

$$Z_m = \left(1 + \frac{2Z_{in}^{*}}{Z_{in}}\right) Z_{out} , \tag{6.35}$$

$$Z_n = \left(1 + \frac{2Z_{in}^{*}}{Z_{in}}\right) Z_{out} . \tag{6.36}$$

Therefore,

$$Z_m = Z_n , \tag{6.37}$$

it proves that the two balanced ports are balanced.

It should be noted that from equation (6.35) or (6.36)

$$Z_m = Z_n = 3Z_{out} \ , \tag{6.38}$$

if Z_{in} is a pure resistor,

$$Z_{in} = R_{in} \ . \tag{6.39}$$

The expression (6.38) shows that Z_m and Z_n are not conjugate- matched to the balanced impedance of two balanced ports respectively. It is not surprising because the LC balun is a network with 3 ports.

Furthermore, we are going to examine the phase relation between the single or input port and the balanced ports. Referring to Figure 6.12,

$$v_1 = \frac{Z_{out} // Z_L}{Z_{out} // Z_L + Z_c} v_o \ , \tag{6.40}$$

$$v_2 = \frac{Z_{out} // Z_c}{Z_{out} // Z_c + Z_L} v_o \ . \tag{6.41}$$

By means of equations (6.14), (6.15), (6.16), (6.17), (6.31) and (6.32), and note that

$$v_o = \frac{Z_{in}^{\ *}}{Z_{in} + Z_{in}^{\ *}} v_{in} \ , \tag{6.42}$$

then,

$$v_1 = j \frac{\sqrt{2\left(R_{out} R_{in} + X_{out} X_{in}\right)}}{4R_{out}} v_i \ , \tag{6.43}$$

$$v_2 = -j \frac{\sqrt{2\left(R_{out} R_{in} + X_{out} X_{in}\right)}}{4R_{out}} v_i \ . \tag{6.44}$$

It means that the voltage at the balanced or output port 1 is 90^o ahead of the single or input port and the voltage at the balanced port 2 is 90^o behind of the single or input port. It is therefore concluded that the two balanced or output ports are differential as long as condition (6.25) is satisfied.

Based on the derivation of all the equations above, one might feel that the design of an LC balun seems to be an easy work. It seems that all we need to do is to calculate the values of L and C from equations (6.31) and (6.32), on the basis of the impedance of single and balanced ports, Z_{in} and Z_{out}. This is, however, not fully true because there is a limitation on the impedance of single and balanced ports, Z_{in} and Z_{out}. From equations (6.31) and (6.32), it is found that the product $Z_{out}Z_{in}^{\ *}$ must be a real number because the value of L or C is real. It implies that

$$R_{in}X_{out} - R_{out}X_{in} = 0 \ , \tag{6.45}$$

or,

$$\frac{R_{in}}{R_{out}} = \frac{X_{in}}{X_{out}} \ . \tag{6.46}$$

It means that the single or input impedance Z_{in} must be proportional to Z_{out} so that the ratio of their real parts is equal to the ratio of their imaginary parts as shown in equation (6.39). As a matter of fact, relationship (6.46) is implied in expressions (6.29) and (6.30) already.

Now, before the calculation for the values of L and C from equations (6.31) and (6.32), we must check out if condition (6.46) is satisfied or not at first. If both the given Z_{in} and the given Z_{out} are pure resistances or reactances, this limitation will automatically disappear. Usually both of the given Z_{in} and Z_{out} are neither pure resistance nor reactance. And if so, Z_{in} and Z_{out} are not compatible with equations (6.31) and (6.32).

The unique solution of the problem is to modify either Z_{in} or Z_{out} by the addition of an extra part, either inductor or capacitor, so that the modified Z_{in}' and Z_{out}' satisfy limitation (6.46). On either Z_{in} or Z_{out} only the imaginary part but not real part is modified so that the modification would not increase unnecessary insertion loss. Furthermore, it is more economic if the modification is at the single portion but not at the differential portion, and is done by a capacitor but not by an inductor.

Let's demonstrate such a modification through an example. Assuming that the original Z_{in} or Z_{out} is

$$Z_{in} = 746.252 + j393.360 \ \Omega, \tag{6.47}$$

$$Z_{out} = 1800.000 - j4500.000 \ \Omega. \tag{6.48}$$

We are going to modify the input or single port impedance Z_{in} by a capacitaor C_m in series, which will change the original value of Z_{in} as shown in (6.47) to the modified value Z_{in}',

$$Z_{in}' = 746.252 - j1865.625 \ \Omega, \tag{6.49}$$

so that the values of Z_{in}' and Z_{out} satisfy limitation (6.46). The value of C_m can be calculated from the difference ΔZ_{in} between Z_{in} or Z_{in}', that is

$$\Delta Z_{in} = Z_{in}' - Z_{in} = (-j1865.625) - j393.360 = -j2258.987 \ \Omega. \tag{6.50}$$

If the operating frequency is, say,

$$f = 73.35\ MHz, \tag{6.51}$$

Then

$$C_m = 0.96\ pF. \tag{6.52}$$

The LC balun is therefore modified from Figure 6.11(a) to Figure 6.13.

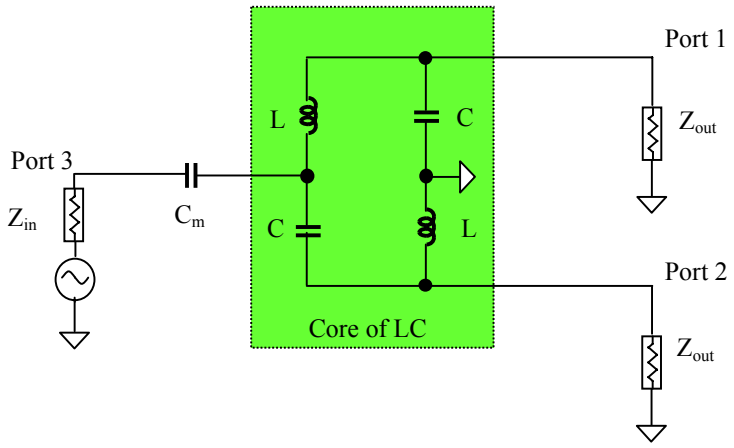


Figure 6.13 Modified LC balun

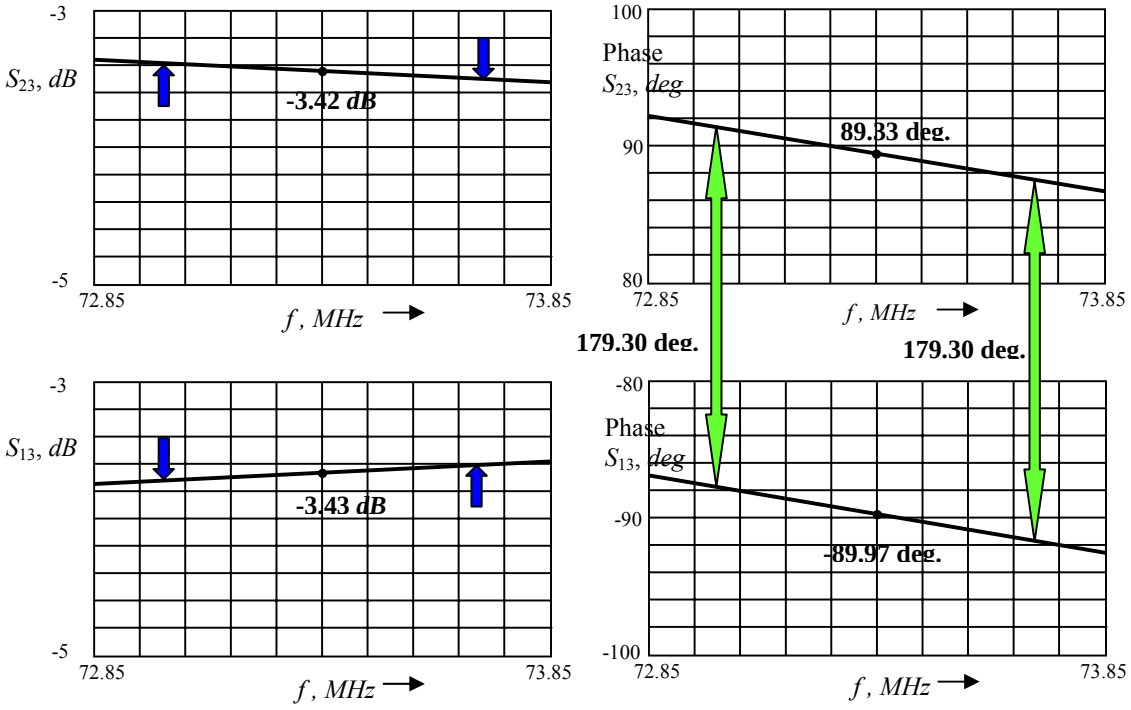


Figure 6.14 Tested results of LC balun built by chip L and C

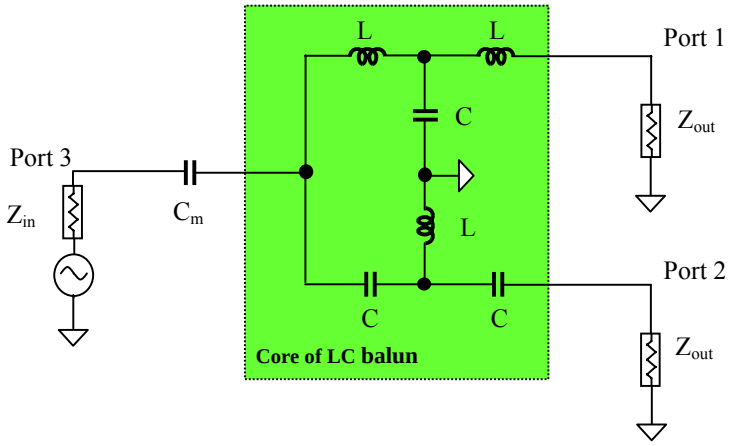
- Terminal 1&2 : differential pair;
- Terminal 3 : single-ended.

When the operating frequency is in the range around 100 MHz, it is easy to build an LC balun by the chip L and chip C parts. Figure 6.14 shows the tested results of an LC balun built by chip inductors and chip capacitors with an operating frequency around 73.35 MHz.

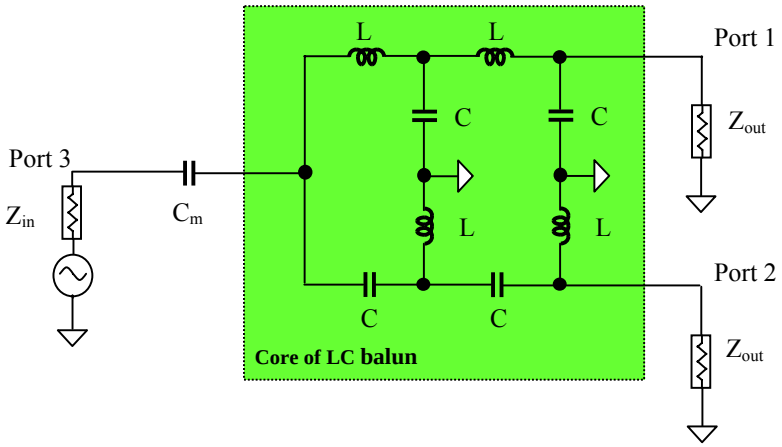
It can be seen that at the central frequency 73.35 MHz, the insertion loss *IL* and the phase θ are

IL = 0.42 dB, $\theta = 89.33^{o}$, for port 2,
IL = 0.43 dB $\theta = -89.97^{o}$, for port 3,

which is quite close to a good differential pair.



(a) Core of LC balun consists of 3 inductors and 3 capacitors



(b) Core of LC balun consists of 4 inductors and 4 capacitors

Figure 6.15 Complicated LC balun

At a lower frequency e.g. 72.85 MHz, the amplitude at port 2 is up and at port 1 is down. The difference is less than 0.2 dB. The phase is simultaneously up at port 1 and port 2.

At a higher frequency e.g. 73.85 MHz, the amplitude at port 2 is down and at port 1 is up. The difference is also less than 0.2 dB. The phase is simultaneously down at port 1 and port 2. The phases at port 1 and 2 are simultaneously up or down. From 72.85 to 73.85 MHz, the phase difference between port 1 and 2 is kept at 179.3^{o}. Consequently, the LC balun is a wideband balun. This should be a surprise to those who thought that the LC balun should be a narrow band balun because the L and C are quite sensitive to the operating frequency.

The LC balun as shown in Figure 6.11 or 6.13 is simplest one. It might be extended to somehow complicated ones, which could have better differential pair with wide band performance but higher insertion loss. Designers must take the trade-off between the advantages and disadvantages. Figure 6.15 (a) shows the core of LC balun constructed by 3 inductors and 3 capacitors and Figure 6.15(b) shows the core of LC balun constructed by 4 inductors and 4 capacitors.

References

[1] J. K. Shimizu; E. M. T. Jones; "A Wide-Band Strip-Line Balun", Microwave Theory and Techniques, IEEE Transaction on, Volume 7, Issue:1, pp. 128-134, January, 1959.

[2] Y. Shimada; "Input impedance Analysis of 1:1 Balun", Microwave Theory and Techniques, IEEE Transaction on, Volume 18, Issue:5, pp. 264-269, May, 1970.

[3] O. M. Woodward; S. M. Perlow; "Balance Quality Measurements on Baluns", Microwave Theory and Techniques, IEEE Transaction on, Volume83, Issue:10, pp. 821-824, October, 1983.

[4] George D. Vendelin; Anthony M. Pavio; Ulrich L. Rohde; "Microwave Circuit Design Using Linear and Nonlinear Techniques" (Book), John Wiley & Sons, Inc., 1990.

[5] S. Lu; A. Elshabini-Riad; S. M. Riad; "Characterization and Modeling of a Wideband Microstrip Balun", Precision Electromagnetic Measurements 1990, CPEM '90 Digest, Conference on, pp. 220-221, June 11-14, 1990.

[6] J. Rogers; R. Bhatia; "A 6 to 20 GHz Planar Balun Using a Wilkinson divider and Lange Couplers", Microwave Symposium Digest, 1991 IEEE MTT-S International, Vol. 2, pp. 865-868, June 10-14, 1991

[7] V. Trifunovic; B. Jokanovic; "New Uniplanar Balun ", Electronic Letters, Volume 27, Issue:10, pp. 519-520, May 9, 1991.

[8] Richard C. Li; "Double Line RF Balun Transformer," U. S. Patent 5,477,204, Motorola Inc., 1993.

[9] M. C. Tsai; "A New Compact Wideband Balun", Microwave and Millimeter Monolithic Circuits Symposium, 1993, Digest of Papers, IEEE 1993, pp. 123-125, June 14-15, 1993.

[10] M. N. Tutt; H. Q. Tserng; A. Ketterson; "A low-Loss, 5.5 GHz-20 GHz Monolithic Balun" , Microwave Symposium Digest, 1997, IEEE MTT-S International, Vol. 2, pp. 933-936, June 8-13, 1997.

[11] T. Rutkowski; W. Zieniutycz; K. Joachimowski; "Wideband Coaxial Balun for Antenna Application", Microwaves and Radar, 1998, MIKON '98, 12 International Conference on, Vol. 2, pp. 389-392, May 20-22,1998.

.

[12] Jwo-Shium Sun; Tsung-Lin Lee; "Design of a planar Balun" , Microwave Conference 2001, APMC 2001, 2001 Asia-Pacific, Vol. 2, pp. 535-538, December 3-6, 2001.

[13] Liang Tao; J. Gillis; D. Wang; P. Cooper; "Design and Modeling of Compact On-chip Transformer/Balun using Multi-Level Metal Windings for RF Integrated Circuits" Radio Frequency Integrated Circuits (RFIC) Symposium 2001, IEEE, May 20-22, 2001.

[14] H. Y. D. Yang; L. Zhang; J. A. Castaneda; "Design and Analysis of a Multi-Layer Transformer Balun for Silicon RF Integrated Circuits" Radio Frequency Integrated Circuits (RFIC) Symposium, 2002, IEEE, June 2-4, 2002.

[15] Jong-Wook Lee; K. J. Webb; "Analysis and Design of Low-Loss Planar Microwave Baluns Having Three Symmetric coupled Lines", Microwave Symposium Digest, 2002 IEEE MTT-S International, Vol. 1, pp. 117-120, June 2-7, 2002.

Index

Contents

Chapter 7 Tolerance Analysis

7.1 Importance of Tolerance Analysis

The development of a product presumably contains three portions: product design, process design, and production. The R&D (Research and Development) department is usually in charge of the first two portions: product and process design, while the manufactory is in charge of the last two portions: process design and production.

When a product has gone through all the test and examination procedures and is going on the production line, people from the R&D and manufactory drink champagne together to celebrate for their successful product in a noble style and a friendly atmosphere. However, there are also times when people from R&D and manufactory scream, shout, and point at each other's nose on account of their failed product: "Who is to blame for the failure of the product?"

To measure success or failure of the product development, the manufacturability of a product is one of key criteria. People in both of R&D department and manufactory must take care of it in all the stages of the development.

A start-up company is doomed to be bankrupt if the expected product is developed in such a way: At the beginning the expected product has been well developed by the R&D department in their laboratory. Following the specifications of the product, the main design steps, including simulation, layout, implementation of module or IC taped-out, the testing and the demonstration, have been accomplished. In these working phases many patents have been rewarded for their creative works, and many papers have been published. However, this expected product has not been analyzed for tolerance in the simulation phase or gone through the ALT (Accelerate Life Testing) in the testing phase. In a hurry the expected product has been put on the production line for mass production. As a result, the yield rate of the expected product is very poor, say, less than 50%. It wastes *more* than 50% of the funds for the direct and indirect material cost, including man-power and the manufactory facility, and the time. The tolerance analysis and the ALT in the developing stage are to ensure the manufacturability of a product. Obviously they are very important steps to guarantee the success for a new product.

Without designing for manufacturability, it might cause a terrible loss of capital or bankruptcy.

Manufacturability of a product means that the product can be reproduced in a massive production line by the features of

1) High yield rate, including the performance of
   - Excess over specification,
     For example, if the specified gain for an LNA design is 10 *dB*, the designed gain should be set on 12 *dB*.

- Identity of product units,
  For example, if the tested gain of an LNA is 12 *dB*, the variation of the gain should be kept in the range of (11 -13) *dB* in the massive production line.
- High reliability.
  All the units of product should be able to pass the ALT.

2) Low cost, including cost of
   - Material and parts,
     The quality of material should be high enough to ensure a 6σ design and production.
     The part count is one of important parameters in the design and on production line. It impacts the product not only at the cost but also at the reliability. The less the parts count, the lower the cost, and the higher the reliability.
   - Manpower,
   - Manufactory maintenance and equipment.

While designing for manufacturability, a designer must take over the responsibilities to

1) To do the integrated product and process designs that allow 6σ defect levels in manufactured units.
2) To provide tolerance to the suppliers so as to ensure 100% usable purchased parts and materials.
3) To suggest a "waste-free" manufacturing system.

Most of the tasks listed above rely on tolerance analysis. It is therefore fitting that in this chapter, we would confine our discussion only to tolerance analysis. The tolerance analysis is the key segment in the engineering design works. In some circuit designs, their performances are mainly limited by the tolerance. The designer must do the tolerance analysis first. The example in section 7.4 is a good evidence.

7.2 Fundamentals of Tolerance Analysis

7.2.1 Tolerance and Normal Distribution

If you purchase a large number of 1 kΩ chip resistors from a company or manufacturer, the ideal condition is that all the units of the bulk product have the same value, that is, all the resistors, or 100% of the resistors, have the same value of resistance, 1 kΩ, as specified for the nominal value. Therefore the deviation of the individual value is zero for all the resistors. In reality, this ideal condition never turns up. The values of the resistors are never exactly 1 kΩ for the whole bulk. The majority of the resistors have values around 1 kΩ, and the number of resistors decreases as the deviation of their value is increased from 1 kΩ. Figure 7.1 plots the relative number of resistors versus the value of resistor.

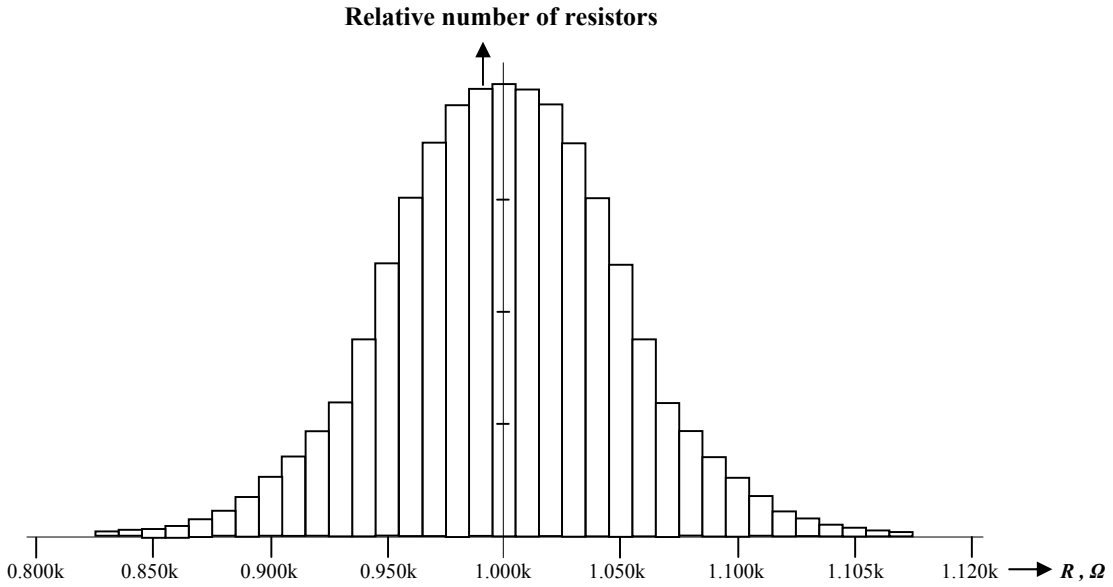


Figure 7.1 Histogram of relative number of resistors versus the value

The histogram is symmetrical to the point, m, so that the value m is its average. It is provided in the specification of the resistor as a nominal value of the resistor, that is,

$$m = 1000\ \Omega\,, \tag{7.1}$$

and the relative tolerance, $Tol_{Relative}$, in respect to the average value, is defined as

$$Tol_{\mathrm{Re}lative} = \frac{\sqrt{\overline{\sum_i \Delta R_i^{\,2}}}}{m} = 5\% \quad , \tag{7.2}$$

where m= Average of resistance of these resistors,
ΔR_i = Deviation of resistance from the average value when R=R_i.

The summation is operated over all the individual values of resistor R_i. The average square root is called variance or standard deviation of these resistors, and is denoted by

$$\sigma = \sqrt{\overline{\sum_i \Delta R_i^{\,2}}} \ \Omega, \tag{7.3}$$

From expression (7.3), expression (7.2) can be re-written as,

$$Tol_{\mathrm{Re}lative} = \frac{\sigma}{m} \ . \tag{7.4}$$

The relative tolerance is simply the ratio of the standard variation to the average value. It reaches zero when σ is or approaches zero, or, when every individual value of resistors is equal to the value m.

Typically the distribution of the resistor's value, R, as shown in Figure 7.1 is a normal probability function with a probability density, $\rho(z)$,

$$\rho(z) = \frac{e^{-\frac{z^2}{2\sigma^2}}}{\sqrt{2\pi}} \quad , \tag{7.5}$$

which is dimensionless and is a function symmetrical to $z = 0$,
where

$$z = \frac{R - m}{\sigma} \quad , \tag{7.6}$$

is also a dimensionless resistance, which is relative to the average value of resistor and is normalized by the variance σ. Figure 7.2 is a plot of $\rho(z)$ versus σ and is overlapped with Figure 7.1. Its well-matched consequence between Figure 7.1 and 7.2 verifies the assertion that the random variable R is a Gaussian or normal probability function. There are two abscissa scales in Figure 7.2. The upper scale, z, is a dimensionless and normalized scale, which corresponds to its normal distribution, while the second scale, R, is the real value of resistor, which corresponds to its Gaussian distribution.

A Gaussian distribution becomes a normal distribution when

$$m = 0 \quad , \tag{7.7}$$

and

$$\sigma = 1 \quad . \tag{7.8}$$

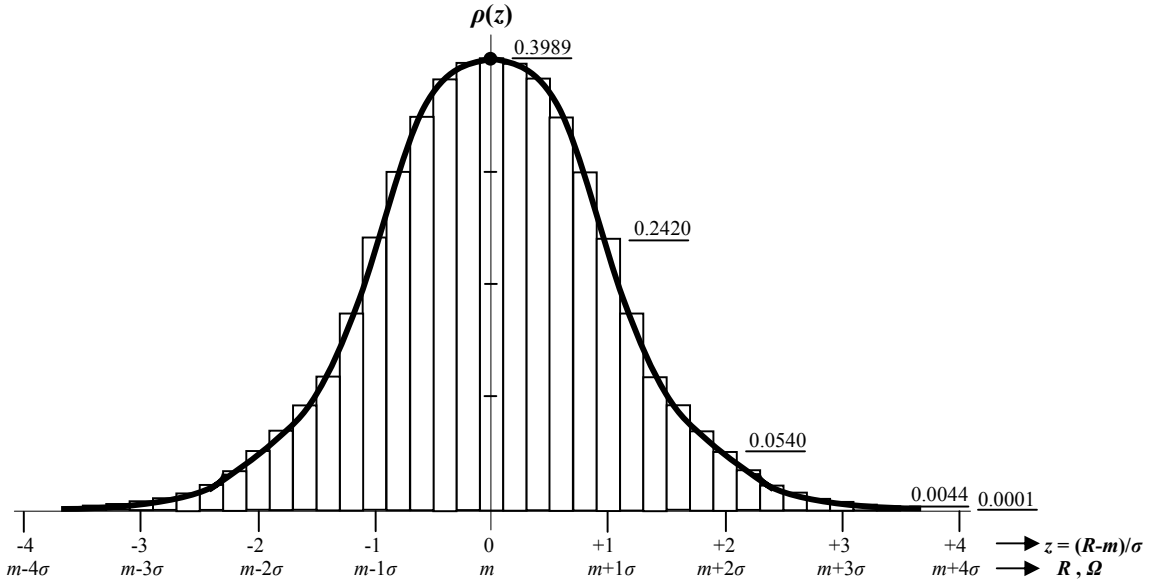


Figure 7.2 Distribution of the random variable, R, is a Normal probability function

It is interesting to see how many resistors there would be when z is varied from 0 to z, or R is varied from m to $m+z\sigma$. This is an integral of the $\rho(z)$, integrated from z=0 to z=z, or from R=m to R= $m+z\sigma$, that is,

$$f(z) = \int_0^z \rho(x)dx = \int_0^z \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi}}dx \quad . \qquad (7.9)$$

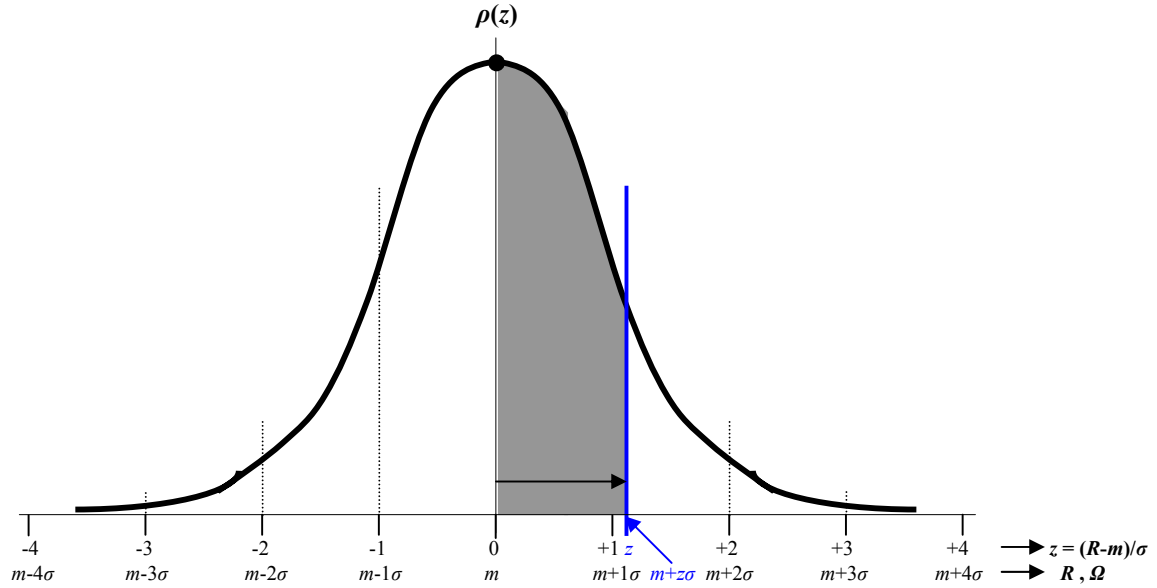


Figure 7.3 Shadowed area, f(z), is an integral of $\rho(z)$ from 0 to z or from m to $m+z\sigma$.

The integral, $f(z)$, is the shadowed area as shown in Figure 7.3.

Normal probability function is related to error function, *erf(x)*, such as,

$$f(z) = \int_0^z \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx = \frac{1}{2} erf\left(\frac{z}{\sqrt{2}}\right) \quad , \qquad (7.10)$$

and

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-y^2} dy \quad . \qquad (7.11)$$

The value of $f(z)$ or the shadowed area in Figure 7.3, depends on the integrated interval along the abscissa. Figure 7.4 shows the values of $f(z)$ in each interval of σ.

It can be seen that

when the interval is	$-1\sigma < z < +1\sigma$,	then the area is	$f(z) = 68.26\%$,
when the interval is	$-2\sigma < z < +2\sigma$,	then the area is	$f(z) = 95.44\%$,
when the interval is	$-3\sigma < z < +3\sigma$,	then the area is	$f(z) = 99.74\%$,
when the interval is	$z < -3\sigma$, and $z > +3\sigma$,	then the area is	$f(z) = 0.26\%$,

In other words if this resistor, m = 1 kΩ and σ = 50 Ω, is applied to a product and is allowed to have a maximum deviation of 6σ, or $-3\sigma < z < +3\sigma$, then, the possibility of the product failure due to this resistor is less than 0.26%.

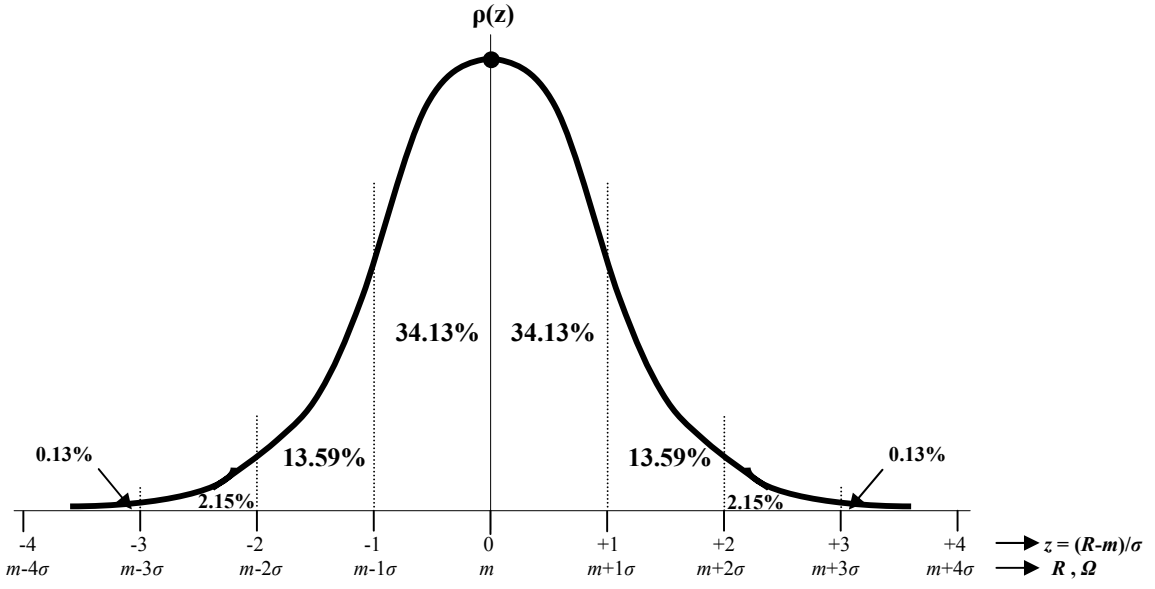


Figure 7.4 At each interval, z, the appearing percentage of a random variable with a normal distribution

7.2.2 6σ, C_p, and C_{pk}

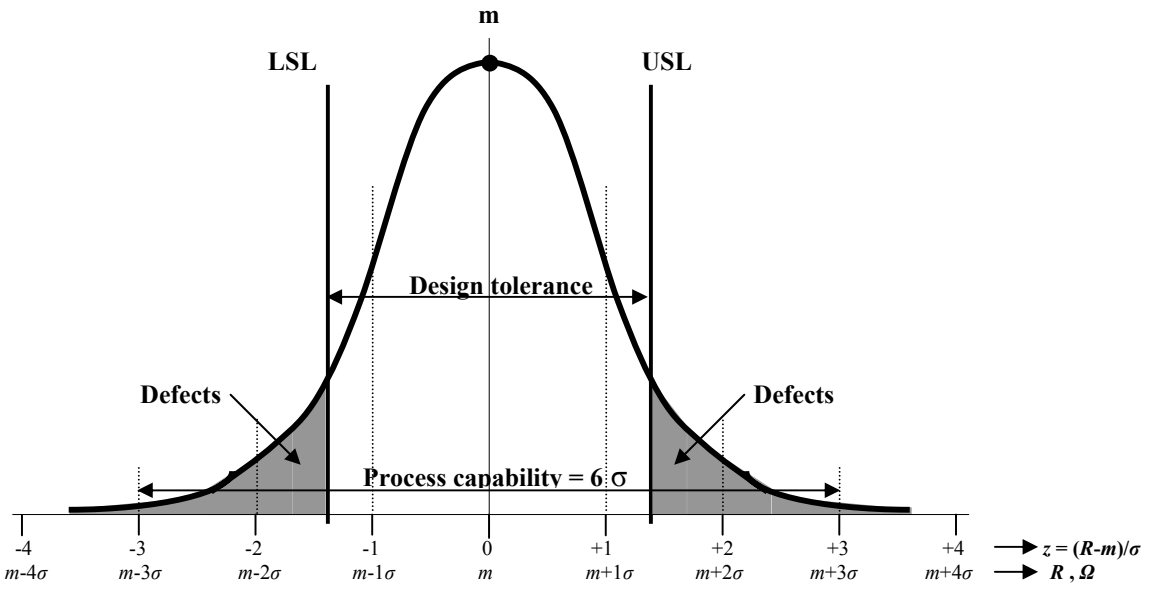


Figure 7.5 The various terminologies of a process with normal distribution

Figure 7.5 indicates some terminologies used in the discussion.

- **Process mean, nominal target, and nominal specification,** are different terminologies but describe the same objective, the **average value of process**, *m*.

- **Process capability** of a part is simply the full range of the normal process variation. Approximately it can be replaced by 6σ because within the interval 6σ, 99.74% of the population is covered.

- **Tolerance of process** is defined as

$$Tolerance = USL - LSL \quad , \tag{7.12}$$

  where USL = Upper specification limit, σ,
  LSL = Lower specification limit, σ.

- **Defects of process** comes from the excessive portions outside the specification limits or the design tolerance.

$$Defects = 1 - f(z = USL) - f(z = LSL) \quad , \tag{7.12}$$

The **Capability Index** C_p is defined as

$$C_p = \frac{Design\_Tolerance}{\Pr ocess\_Capability} \approx \frac{USL - LSL}{6\sigma} \quad , \tag{7.14}$$

in order to quantitatively describe the relationship between the design tolerance and the process capability. A high *Cp* index indicates that the process is capable of reproducing the characteristic we expected. The numerator in expression (7.14) represents the design capability of a designer who applies this part. The denominator in expression (7.14) represents how far the part's characteristic deviates from its mean, which is expected to be as small as possible. The excellent design is that the design tolerance is allowed to cover total process variation or process capability, that is,

$$C_p = 1 \quad . \tag{7.15}$$

In order to quantitatively describe the deviation of the actual process mean from the nominal mean, another index, **Adjusted Capability Index**, C_{pk}, is defined as

$$C_{pk} = C_p(1 - K) \quad , \tag{7.16}$$

where

$$K = \frac{m - m'}{(USL - LSL)/2} \quad , \tag{7.17}$$

where m = Nominal process mean, σ,
m' = Actual process mean, σ.

The optimum condition is

$$C_{pk} = C_p \ , \tag{7.18}$$

when

$$K = 0 \ , \text{ or } \quad m = m' \ . \tag{7.19}$$

The following 4 examples, as shown in Figure 7.6 to Figure 7.9, might be helpful to understand the relationship between the capability index, C_p, adjusted capability index, C_{pk}, and other parameters. The first 2 examples are the cases that the process mean is equal to the specified mean, m=m'. It implies that the process has been well controlled. The defects rely on the design tolerance. In first example the defects, 5.56%, are higher than that in second example, 0.26%, because in first example the design tolerance is 4σ while in second example it is 6σ. In the manufacturing field, 0.26% of defects is equal to an "OK" rate, 99.74%. Quantitatively we might define the "OK" rate as

$$OK\_rate = 1 - Defects \quad . \tag{7.20}$$

The second example, as a matter of fact, is quite close to be a perfect or an ideal condition, in which the defects are 0% or the "OK" rate is 100%. A 6σ production line

means that every part has been designed with 6σ tolerance, and their processes have been stringently controlled. Mathematically, a 6σ design includes

$$USL - LSL = 6\sigma \ , \quad \text{or} \quad C_p = 1 \ , \tag{7.21}$$

$$m = m' \ , \quad \text{or} \quad C_{pk} = C_p = 1 \ . \tag{7.22}$$

The 1st condition, (7.21) indicates that a large deviation is allowed in the design of the product, or, on other words, the yield rate of product is approached to almost 100% within the specified range of goals. The 2nd condition indicates that the processing is well controlled so that the average value is exactly the same as what is expected.

In the1980's, Japanese manufacturers initialized a productive movement called "6σ design and production". Today, 6σ design and production becomes the main technical goal of product, which is pursued by enterprises and companies over the world.

Theoretically, the defects can be further reduced down if the tolerance of design and production is more than 6σ. However, the further reduction of defects is tiny because by 6σ design or production the defects are only 0.26%. To pursue the tolerance of design or production more than 6σ, a tremendous effort must be put on whereas only a tiny benefit can be obtained. This is why people keep the goal of the process tolerance on the "6σ design and production".

Example #1

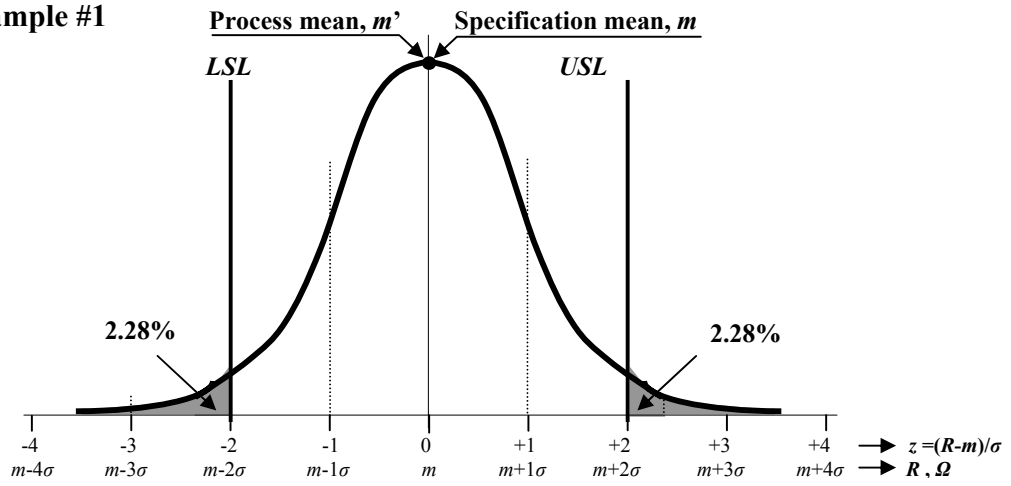


Figure 7.6 A normal distribution with m=10, σ =0.01, USL-LSL=4σ, m'=m .

$$K = \frac{\left| m - m' \right|}{(USL - LSL)/2} = 0$$

$$Defects = 1 - 2f(z = 2\sigma) = 1 - 2 * 47.72\% = 5.56\%$$

$$C_p = \frac{USL - LSL}{6\sigma} = \frac{4\sigma}{6\sigma} = 0.6667$$

$$C_{pk} = C_p(1 - K) = 0.6667(1 - 0) = 0.6667$$

Example #2

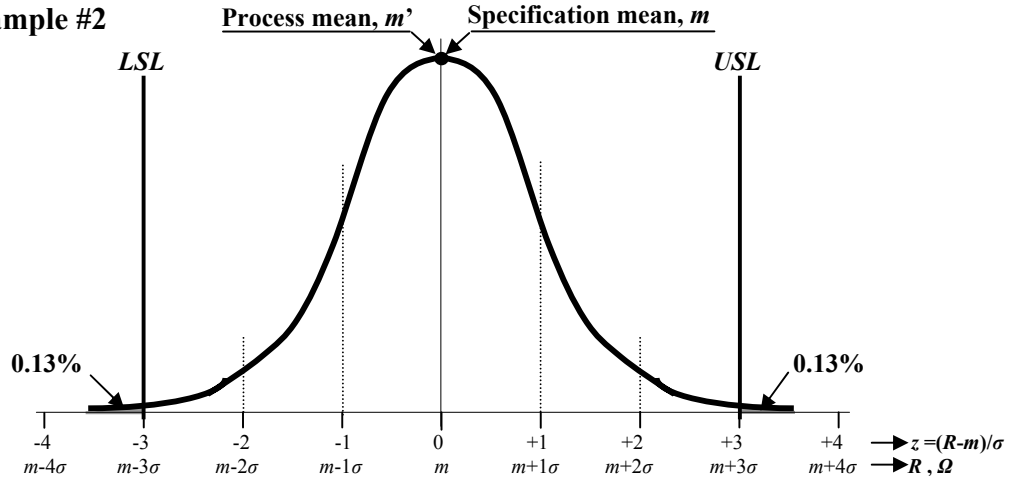


Figure 7. 7 A normal distribution with m=10, σ =0.01, USL-LSL=6σ, m'=m .

$$K = \frac{\left|m - m^{'}\right|}{(USL - LSL)/2} = 0$$

$$Defects = 1 - 2f(z = 3\sigma) = 1 - 2 * 49.87\% = 0.26\%$$

$$C_p = \frac{USL - LSL}{6\sigma} = \frac{6\sigma}{6\sigma} = 1.0000$$

$$C_{pk} = C_p(1 - K) = 1.0000(1 - 0) = 1.0000$$

Example #3

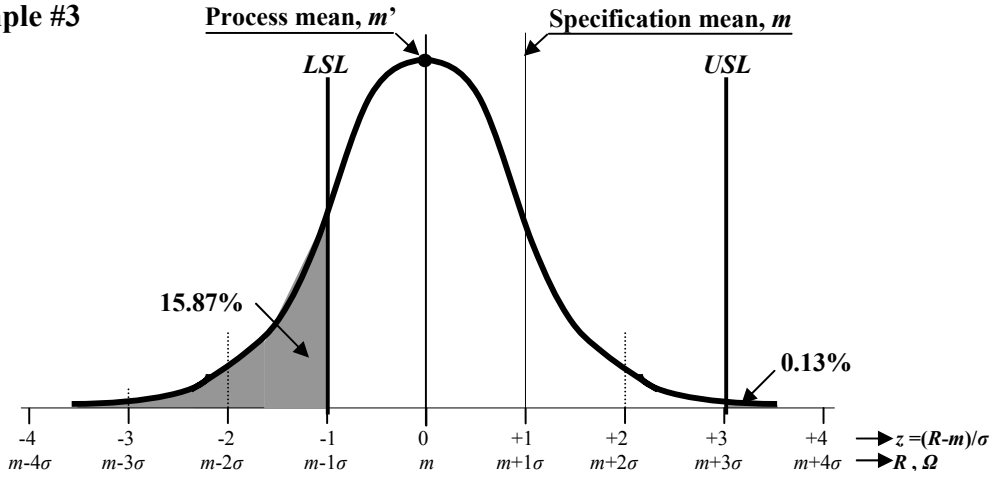


Figure 7.8 A normal distribution with *m*=10, *σ* =0.01, *USL-LSL*=4*σ*, *m*'=9.99 .

$$K = \frac{\left|m - m^{'}\right|}{(USL - LSL)/2} = \frac{1}{4/2} = 0.5000$$

$$Defects = 1 - f(z = 3\sigma) - f(z = -\sigma) = 1 - 34.13\% - 49.87\% = 16\%$$

$$C_p = \frac{USL - LSL}{6\sigma} = \frac{4\sigma}{6\sigma} = 0.6667$$

$$C_{pk} = C_p(1 - K) = 0.6667(1 - 0.5) = 0.3334$$

Example #4

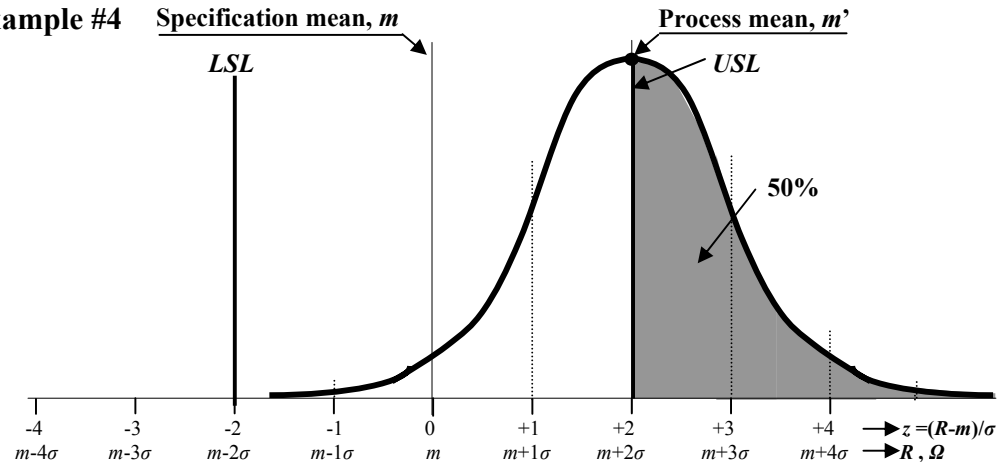


Figure 7. 9 A normal distribution with m =10, σ = 0.01, USL-LSL = 4σ,

$$K = \frac{\left|m - m'\right|}{(USL - LSL)/2} = \frac{2}{4/2} = 1$$

$$Defects = 1 - f(z = 0\sigma) - f(z = -4\sigma) = 1 - 0\% - 50\% = 50\%$$

$$C_p = \frac{USL - LSL}{6\sigma} = \frac{4\sigma}{6\sigma} = 0.6667$$

$$C_{pk} = C_p(1 - K) = 1.0000(1 - 1) = 0$$

The third and fourth example are the cases when the process is not well controlled, in which the process mean is not equal to the specification mean. In these two examples, the difference is only the deviation of the mean difference between the process and the specification, K, or m-m' while the other parameters, m, σ, and USL-LSL are the same. It can be seen that the defects, 16%, in example 3 is less than that, 50%, in example 4. This is due to that in the example 4, m-m' is 2σ while in the example 3 it is 1σ.

7.2.3 Yield Rate and DPU

There are many yield rates that have been defined in different fields, such as,

- Process yield is the ratio of the completed units to the started units,
- Test or inspection yield rate is the ratio of the passed units to the tested or inspected units,
- *FTY* (First time yield rate) is the ratio of the first time passed units to the first time tested or inspected units,
- Rolled yield rate is the ratio of the units going through the entire process without any defect to the processed units.

It should be noted that the process yield rate is quite different between the repairable and un-repairable products. For repairable products, the process yield rate is virtually 100%. The number of completed units is equal to the number of the started units because it is repairable. For un-repairable products, the process yield rate might be significantly lower than 100% because many completed units might be declared as unusable or in un-normal status and yet cannot be fixed.

Also, the rolled yield rate is quite different between the repairable and non-repairable products. For repairable products, the rolled yield rate might be significantly lower than 100% of the process yield rate because the process yield rate of repairable product is virtually 100% though many units might be waiting for repair. For un-repairable product, the rolled yield rate is the same as process yield rate if the units going through the entire process are defect-free.

A better parameter to measure the yield rate or the quality of product process is *DPU* (Defects Per Unit), which accounts all defects produced in the process, regardless of product types. The defects per unit, *DPU*, is defined as

$$DPU = \frac{N_d}{N_p} \quad , \tag{7.23}$$

where N_d = Number of defects found at all acceptance points,
N_p = Number of units processed.

As an example, let's consider the probability that a unit of product contains no defects, $P\{0\}$. Assuming that

- This product is manufactured in 10 divisions,
- The "equal opportunity" for creating a defect in these 10 divisions and the *DPU* in every division is

$$DPU_i = \frac{1}{10} \quad , \tag{7.24}$$

where i = 1,2,3……10.
On the average, there is 1 defect per unit in all 10 divisions:

$$DPU = 1, \tag{7.25}$$

The probability that none of 10 divisions contains a defect is a joint probability which combines the probabilities from all the divisions, such as

$$P\{0\} = \prod_{1}^{10} P_i\{0\} \ , \tag{7.24}$$

where $P\{0\}$ = Joint probability,
$Pi\{0\}$ = Defect-free probability in i^{th} division.

From (7.24),

$$P_i\{0\} = 1 - DPU_i = 90\% \ , \tag{7.27}$$

Then, expression (7.24) becomes

$$P\{0\} = 0.9^{10} = 0.348678 \ , \tag{7.28}$$

The probability P{0} is declined when the number of divisions is increased.

$$P\{0\} = 0.9^{100000} = 0.347878 \ , \tag{7.29}$$

when the number of divisions is 100000. It is approaching a limit,

$$P\{0\} \to \frac{1}{e} = 0.347879 \ , \tag{7.30}$$

As a matter of fact, this example is a special case of a Poisson distribution.

7.2.4 Poisson Distribution

By Poisson distribution, the occurrence of random defects of a manufactured product can be statistically predicted.

Poisson distribution is a mathematic model to describe the probability distribution of the arrived entities. For instance, in one hour lunch time, say, statistically from 12:00 to 13:00 o'clock, the number of customers getting into a restaurant for lunch. Assuming that the average number of customers is μ, and the random number of arrivals at the same time interval is x, then the probability of the random arrival number, $P\{x\}$, is

$$P\{x\} = \frac{\mu^x e^{-\mu}}{x!} \ , \tag{7.31}$$

Now, let's return to the example in the previous section where "space" of divisions rather than "time" of 1 hour lunch time characterizes the probability event, and where "1 defect per unit in all of divisions on average" rather than "the average number of customers, μ " is the state of event. Therefore, we have

$$\mu = 1 \ , \tag{7.32}$$

On the other hand, the probability of defect-free means that

$$x = 0 \ . \tag{7.33}$$

By substituting expressions (7.32) and (7.33) into (7.31), the probability of defect-free in all the divisions is

$$P\{0\} = \frac{1^0 e^{-1}}{0!} = \frac{1}{e} , \tag{7.34}$$

which is the answer as shown in expression (7.30).

The example in previous section is a special case. As matter of fact, the probability of defect-free can be obtained by Poisson distribution if its random variable, x, represents DPU rather than just a special case, DPU=1, that is,

$$P\{0\} = \frac{DPU^0 e^{-DPU}}{0!} = e^{-DPU} , \tag{7.35}$$

this is the relationship between the defect-free probability and the average DPU. The first time yield, FTY, can be approximated by the formula:

$$FTY = e^{-DPU} , \tag{7.36}$$

And, in general,

$$P\{x\} = \frac{DPU^x e^{-DPU}}{x!} , \tag{7.37}$$

is the probability for x defects.

In a circuit design, the DPU tends to be proportional to the parts count. Therefore, the ratio of DPU and parts count seems to be a better parameter to compare the quality between different circuit designs. This new ratio is called Defect rate by the unit of PPM/part, or sometimes is just called PPM/part, that is,

$$Defect\_rate_{PPM/part} = PPM/part = \frac{DPU}{Parts\_Count} . \tag{7.38}$$

If a process with N steps, 1, 2, 3, 4,…N, and the value of DPU in each step is $a, b, c, d,...n,$ respectively. Then, in terms of formula (7.36), the first time yield, FTY, is e^{-a}, e^{-b}, e^{-c}, e^{-d}...e^{-n} correspondingly. The fist time rolled yield for the process is

$$FTY_{rolled} = FTY\{A\} * FTY\{B\} * FTY\{C\} * FTY\{D\}..... * FTY\{N\} = e^{-(a+b+c+d+.....+n)} . \tag{7.39}$$

Regardless of process flow or order, the rolled yield can be calculated from the summation of the DPU values in all the steps.

Let's apply definition (7.38) and equation (7.39) to calculate the rolled yield for a handset of cellular phone. Assuming that there are totally 4 steps: parts checking, parts placed, parts soldered, and parts assembly.

a) Defect rate in parts checking = 250 *PPM*,
   it implies that 250 parts fail in 1million parts.
b) Defect rate in parts placement = 150 *PPM*,
   it implies that 150 parts fail in 1million parts.
c) Solder defect rate = 200 *PPM*,
   it implies that 200 parts fail in 1million parts.
   On average, there are 3 connections for each part.
   Therefore, the connection fail rate = 600 PPM.
d) Defect rate in parts assembly = 50 *PPM*,
   it implies that 50 assembly-components fail in 1 million parts.

Table 7.1 lists the calculated *FTY* for a board assembly containing 50, 100, 200, 300 parts.

Table 7.1 Calculated *FTY* for a printed circuit board assembly by 3 process steps

# of parts (Part count)	Parts Checking *a*	Parts Placement *b*	Parts Soldered *c*	Parts Assembly d	Total $a+b+c+d$	Rolled FTY $e^{-(a+b+c+d)}$
50	0.0125	0.0075	0.030	0.0025	0.0525	94.9%
100	0.025	0.015	0.060	0.005	0.105	90.0%
200	0.050	0.030	0.120	0.010	0.210	81.1%
300	0.075	0.045	0.180	0.015	0.315	73.0%

From Table 7.1 it can be seen that the number of parts or part count is an important parameters in a production line. The more part count for a product the lower the rolled FTY!

7.3 An Approach to 6σ Design and Production

From the manufacturability viewpoint, the Capability index C_p is expected to be as high as possible. As shown in expression (7.13), the maximum of C_p is 1, or in other words the designed tolerance is equal to 6σ.

However, from the circuit design viewpoint, the expectation of 6σ design is a tough task to a design engineer. It implies that the circuit block must be in normal performance when not only every step of process but also all the values of the parts adapted, such as resistors, capacitors, inductors, transistors, and so on, are varied from (m-3σ) to (m+3σ), where m is the specified mean or nominal value of the part. The "normal performance" means that the design goals for the circuit must be reached or over. For instance, in the 6σ design for an LNA, the design goal of the power gain, noise figure, *IP*3, and so on, must be reached or over even when all the resistors, capacitors, inductors, transistors, and so on, are varied 6σ around their nominal value or mean.

As mentioned above, the failure rate of product is less than 0.26% or the yield rate of product is higher than 99.74% when it is a 6σ design. At present, the design level is below 6σ in most R&D works over the world.

It is possible to improve the design quality by means of changing the component's σ value. Figure 7.10 illustrates such an idea. Originally the design quality is depicted in the upper portion of Figure 7.10. It is a 3σ design or production with specified *USL* and *LSL* values and the part's σ value. Obviously it is not a perfect design and the defects of product are 2•6.68%=13.36%. With the same specified values of *USL* and *LSL*, the original 3σ design or production in the upper portion of Figure 7.10 can be improved and converted to 6σ design or production as shown in the bottom portion of Figure 5.10. Such a conversion can be realized by the replacing original parts by new parts if the σ value of new parts is lower that the σ value of original parts. Consequently, the original defects of 6.68% would be reduced to 0.26%.

In order to convert 3σ to 6σ design or production, there are two tasks to be done:

1) With the replacement of old parts by new parts, the σ value is changed from the original value to its half,
2) With the replacement of old parts by new parts, the mean, m, is kept unchanged.

For example, if a circuit block consists of one resistor, one capacitor, and one inductor, and their m and σ value before and after replacement are listed in Table 7.2.

Table 7.2 Replacement of original resistor, capacitor and inductor by new ones

	Original value before replacement		New value after replacement	
	m	σ	m	σ .
Resistor	1 k*Ω*	**100 *Ω***	1 kΩ	**50 *Ω***
Capacitor	5 *pF*	**0.4 *pF***	5 *pF*	**0.2 *pF***
Inductor	5 *nH*	**0.3 *nH***	5 *nH*	**0.15 *nH***

With the same specified values of *USL* and *LSL*, if the original is a 3σ design, then it is improved to 6σ design after replacement. The price is a higher cost. In the discrete RF circuit design, the cost of the part with lower σ value is more expensive than that of the part with high σ value. In the RFIC design the configuration of the part with lower σ value is more complicated than that of the part with higher σ value. And, usually the size of the part with lower σ value is much larger than that of the part with higher σ value.

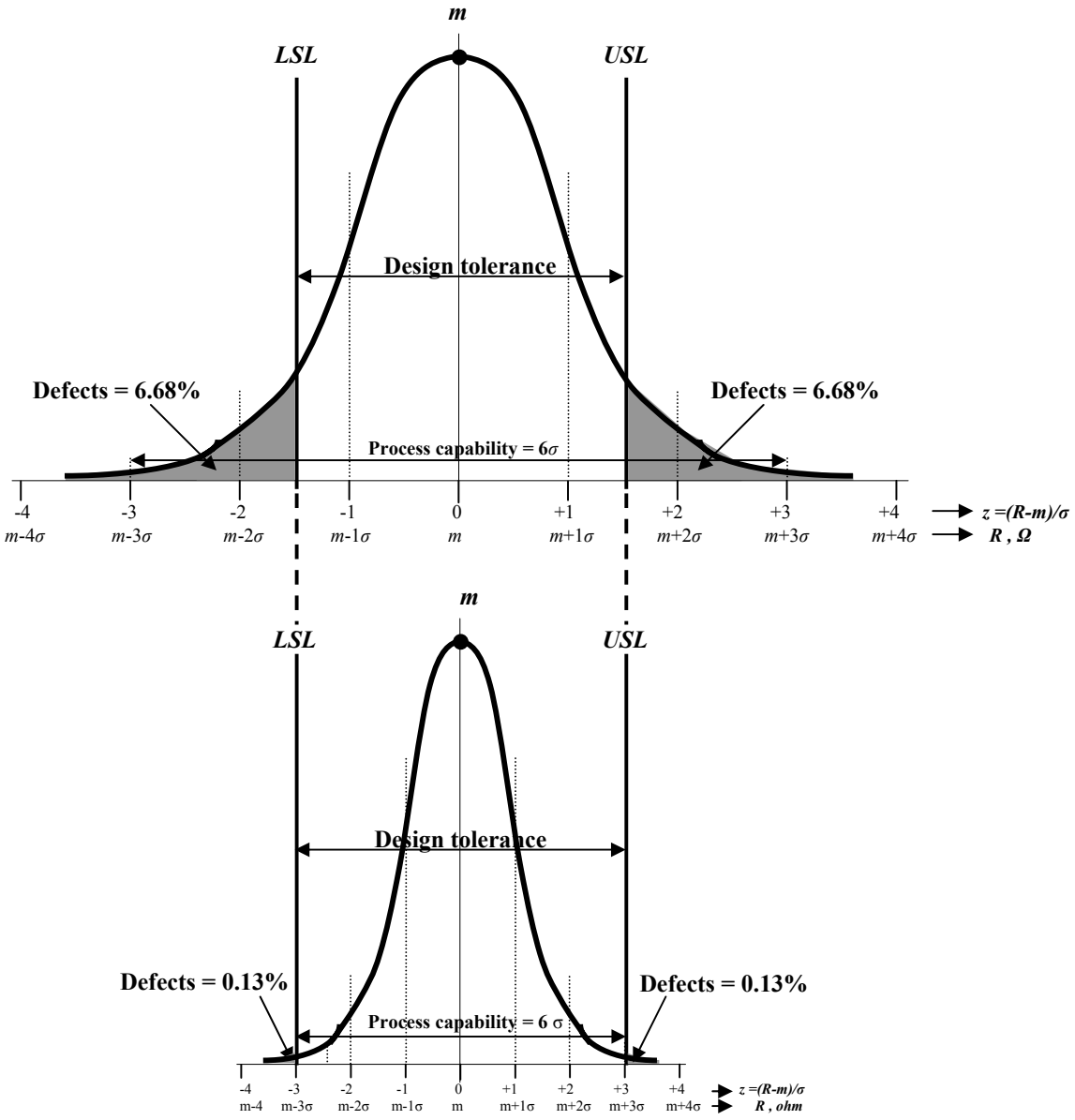
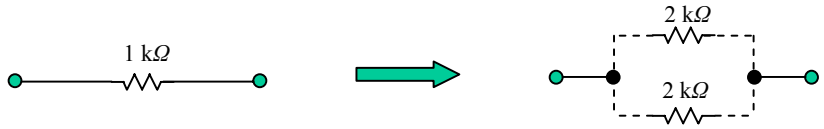


Figure 7.10 Difference of defects when the value of σ is 50% decreased

There is an alternative way to improve the design quality by tolerance. In the real parts, such as resistors, capacitors, and inductors, the σ value is decreased as the value of parts increased if they are produced with the same processing. By utilizing this feature of parts, it is very helpful to approach the 6σ design either for the discrete RF module or RFIC circuit. Let's see how it works for the example that we just discussed. The 1 kΩ resistor would be replaced by two 2 kΩ resistors in parallel and 5 *pF* capacitor would be replaced

Table 7.3 Replacement of original 1 resistor, 1 capacitor, and 1 inductor by 2 resistors, 2 capacitors, and 2 inductors.

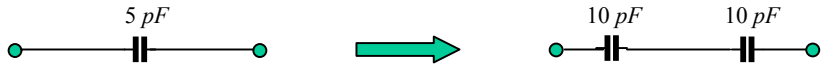
	Original value before replacement		New value after replacement	
	m	σ	*m*	σ .
Resistor	1 kΩ (1 resistor)	**100 Ω**	1 kΩ (2 resistors in parallel)	**50 Ω**
Capacitor	5 *pF* (1 capacitor)	**0.4 pF**	5 *pF* (2 capacitors in series)	**0.2 *pF***
Inductor	5 *nH* (1 inductor)	**0.3 *nH***	5 *nH* (2 inductors in parallel)	**0.15 *nH***



m = 1k**Ω**
σ = 100 Ω = 10% of *m*

m = 2 k**Ω/2 = 1** k**Ω**
σ = 50 Ω = 5% of *m*

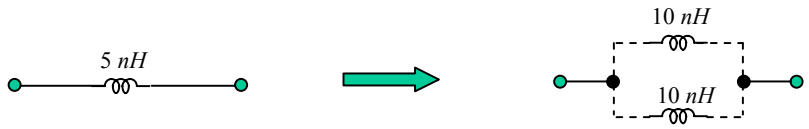
(a) One 1 kΩ resistor is replaced by two 2 kΩ resistors in parallel



m = 5 *pF*
σ = 0.4 *pF* = 8% of *m*

m = 10 *pF*/**2 = 5** *pF*
σ = 0.2 *pF* = 4% of *m*

(b) One 5 *pF* capacitor is replaced by two 10 *pF* capacitors in series



m = 5 *nH*
σ = 0.3 *nH* = 6% of *m*

m = 10 *nH*/**2 = 5** *nH*
σ = 0.15 *nH* = 3% of *m*

(c) One 5 *nH* inductor is replaced by two 10 *nH* inductors in parallel.

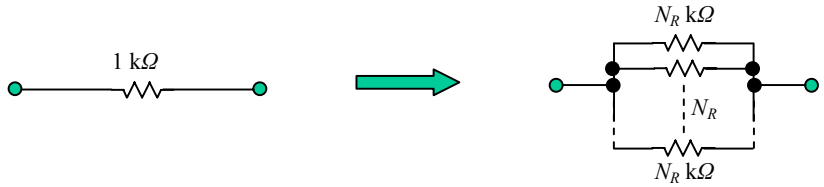
Figure 7.11 One part is replaced by two parts.

by two 10 *pF* capacitors in series. A 3σ design can be converted to a 6σ design if their *m* and σ values are as those listed in Table 7.3 and the replacement is made as shown in Figure 7.11.

In reality it is not necessary to be confined on the replacement one part by two parts. One can perform the replacement by different combinations as shown in Table 7.4 and Figure 7.12.

Table 7.4 Replacement of original resistor, capacitor, and inductor by N_R resistors, N_C capacitors, and N_L inductors.

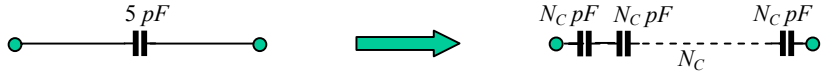
	Original value before replacement		New value after replacement	
	m	σ	*m*	σ .
Resistor	1 kΩ (One resistor)	**100 Ω**	1 kΩ (N_R resistors in parallel)	**50/N_R Ω**
Capacitor	5 *pF* (One capacitor)	**0.4 *pF***	5 *pF* (N_C capacitors in series)	**0.4/N_C *pF***
Inductor	5 *nH* (One inductor)	**0.3 *nH***	5 *nH* (N_L inductors in parallel)	**0.3/N_L *nH***



m = 1k**Ω**
σ = 100Ω = 10% of *m*

m = **1** k**Ω**
σ = 50Ω = 5% of *m*

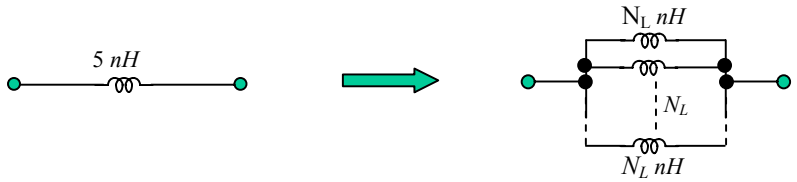
(a) One resistor of 1 kΩ is replaced by N_R resistors of N_R kΩ in parallel



m = 5 *pF*
σ = 0.4 *pF* = 8% of *m*

m = **5** *pF*
σ = 0.2 *pF* = 4% of *m*

(b) One capacitor of 5 *pF* is replaced by N_C capacitors of N_C *pF* in series



m = 5 *nH*
σ = 0.3 *nH* = 6% of *m*

m = **5** *nH*
σ = 0.15 *nH* = 3% of *m*

(c) One inductor of 5 *nH* is replaced by N_L inductors of N_L *nH* in parallel.

Figure 7.12 One part is replaced by the combination of parts.

If a circuit contains more than 1 part, the replacement by more parts with higher part's value but lower variance relatively will certainly improve the design quality so as to approach the 6σ design or production level. The price is more parts count, more area, and more cost to be paid. Here a question might be raised up: Does every part have the same importance in the approach to the 6σ design or production? The answer is No! The importance of each part is not the same. It depends on its location in the topology of circuit, on the current drain, and other factors. We will discuss this partially in the following section.

7.4 An Example: A Tunable Filter Design

By the analysis of this example, we can see how to ensure 100% yield rate in a design, how to control the tolerance of parts, and how to affect the performance from each part.

A quite special feature of this design, the design of a tunable filter, is that the main goal of performance depends on the tolerance analysis. Consequently, it is distinguished from other block designs in that the designer must do the regular simulation for the circuit on one hand and do the tolerance analysis on the other hand, and then take a trade-off between the optimized simulation value and the value from the tolerance analysis for all the parts.

7.4.1 Description of the Tunable Filter Design

The tunable filter is sometimes called the tracking filter.

This tunable filter design is for the receiver front end of a *UHF* portable radio. The central frequency must be able to tune in the entire *UHF* radio band, from 403 to 520 *MHz*. On the other hand, the ideal bandwidth is an audio channel, say, 25 k*Hz*, for the radio-call performance. However, such a narrow bandwidth could be done only by the crystal filter, which is quite expensive and is somehow difficult technically. In light of the trade-off consideration of technology and the cost, the tunable filter in the receiver front end for a portable radio or a cellular phone is implemented by chip capacitors, chip inductors, chip resistors, and the discrete varactors. At *UHF* range, its minimum of bandwidth is usually more than 10 *MHz*, which is far from our expectation. Nevertheless it is still very helpful to improve the selectivity of the receiver front end in the reduction of noise level and the prevention of a variety of spurious interference.

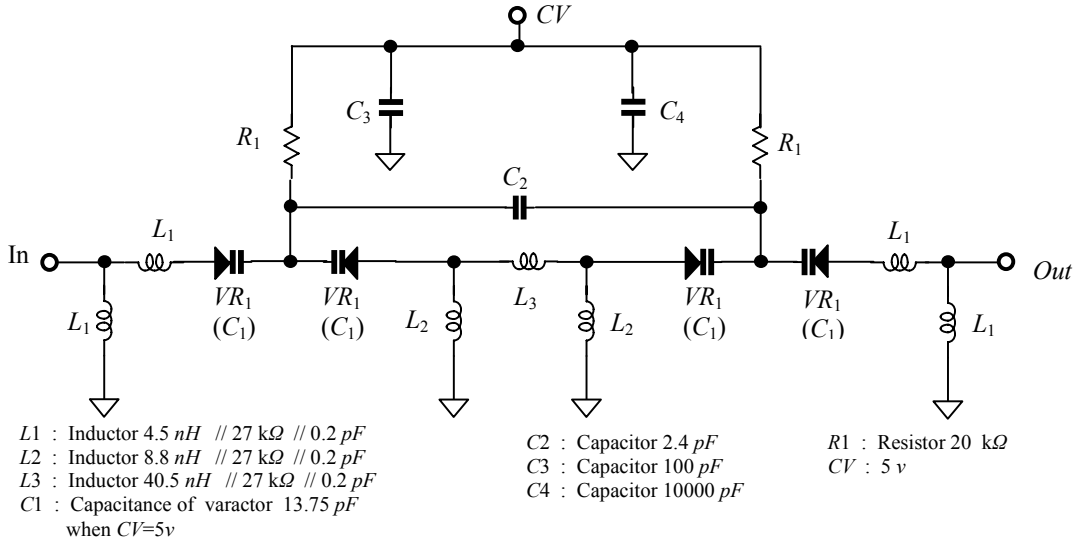


Figure 7.13 Schematic of a *UHF* tunable filter

Figure 7.13 shows the schematic of a *UHF* tunable filter.

The main feature of the circuit is two identical tank circuits coupled by an inductor. The tank circuit consists of the inductors L_1 and L_2, and varactors VR_1. The coupling inductor is L_3. In order to match the impedance of input or output, 50 Ω, two identical inductors, $2xL_1$, in each tank are specially arranged. There is second coupling capacitor, C_2, which creates two "zeros" in the skirt portion of the frequency response plot. One is located below the pass-band and another is above pass-band. These two "zeros" can be adjusted and traced by C_2, L_2, L_3, and VR_1, and can be applied to trace the imaginary spurious in both cases of either high side or low side injection, if the receiver is not operated with direct conversion mode. In the pass-band, the effect of the second coupling is negligible. The tuning function is performed by the 4 identical varactors which are controlled by the control voltage, CV, via the resistors, R_1.

The theory about the coupling of two tank circuits was presented several decades ago. It seems redundant to repeat the basic analysis here. However, the differences of this design from previous are two aspects:

1) In the previous designs, the bandwidth usually is changed from narrow to un-acceptable wide when the central frequency is adjusted from low end to high end of the *UHF* band. The bandwidth is kept almost constant when the central frequency is adjusted within the entire *UHF* band. This is the main reason by which a US patent was offered to this design. The 1st coupling or the main coupling is emphasized to apply an inductor rather than a capacitor or a combination of inductors and capacitors.

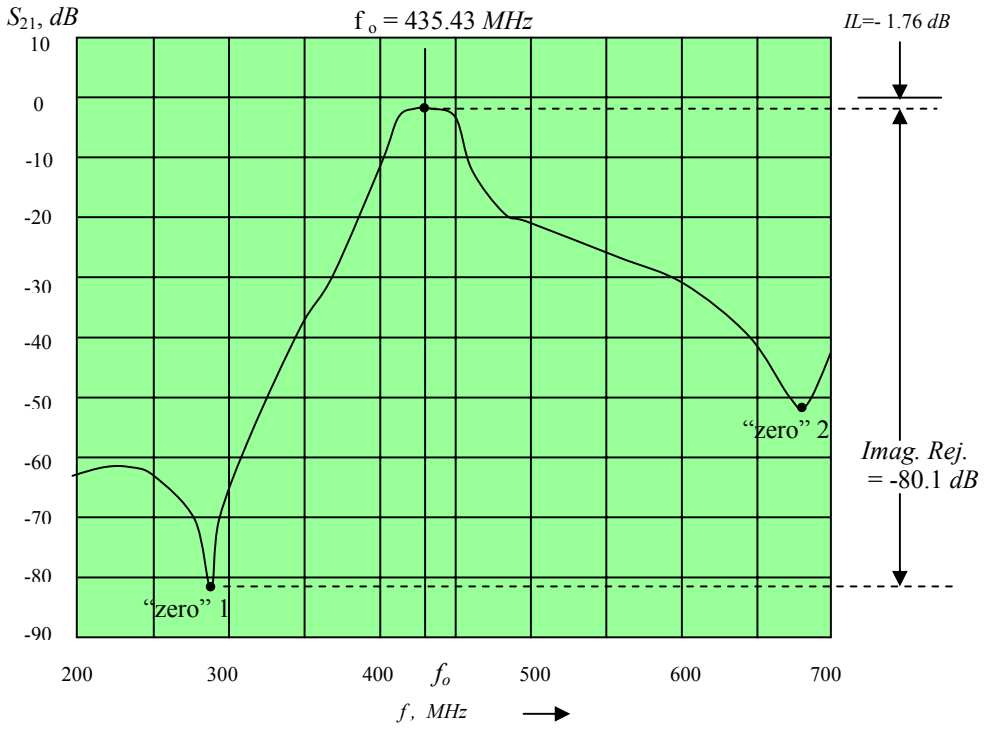


Figure 7.14 Typical frequency response of tunable filter without tolerance

2) The second coupling creates two tractable "zeros" which greatly enhances the image rejection capability for a non-direct conversion receiver.

A typical frequency response of this tunable filter is depicted in Figure 7.14. It shows that

- Central frequency, f_o = 435.43 *MHz*,
- Bandwidth, BW = 33.5 *MHz*,
- Insertion loss, IL = -1.76 *dB*,
- Image Rejection 1, or "zero" 1, $IMRJ_1$ = -80.1 *dB*,
- Image Rejection 2, or "zero" 2, $IMRJ_2$ = -52.4 *dB*.

Figure 7.14 shows the test results of the tunable filter for the low side injection radio. The deep rejection is therefore emphasized at the low side below pass-band, at the "zero" 1. As a matter of fact, it could be emphasized at the high side above pass-band if the tunable filter is designed for the high side injection radio.

As mentioned above, the bandwidth should be as narrow as possible. The specialty of this design is that the bandwidth must be determined from the tolerance of the parts. The tolerance of the parts sets the minimum of the bandwidth.

7.4.2 Monte-Carlo Analysis

The tolerance evaluation of a circuit can be conducted by the Monte-Carlo analysis. The simulation can be set up as shown in Figure 7.15.

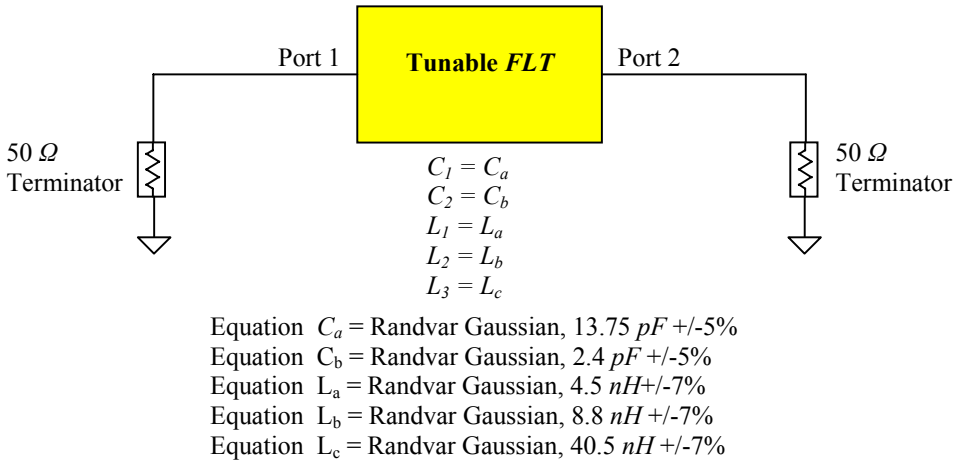


Figure 7.15 Simulation page for Monte-Carlo analysis
(Strictly speaking, it is not a Gaussian but a normal distribution here.)

The purpose of the Monte-Carlo analysis is to find the effect of the parts' tolerance on the performance of the tunable filter. In Figure 7.15 there are two capacitors and 3 inductors involved in the tolerance analysis. They are random variables with Gaussian distribution around their own nominal value or mean, m, and tolerance by percentage, σ, as shown in

Figure 7.15. The nominal value or mean, m, and tolerance by percentage, σ, of parts are provided by the manufacturers. In the Monte-Carlo analysis, the iteration number in sampling of the random variables is set large enough so that the simulated results reflect all the variations. In order to have a clear show, Figure 7.16 contains only 11 iterations, which contains 11 curves of frequency response. In a practical simulation, the number of iterations is set 50 at least.

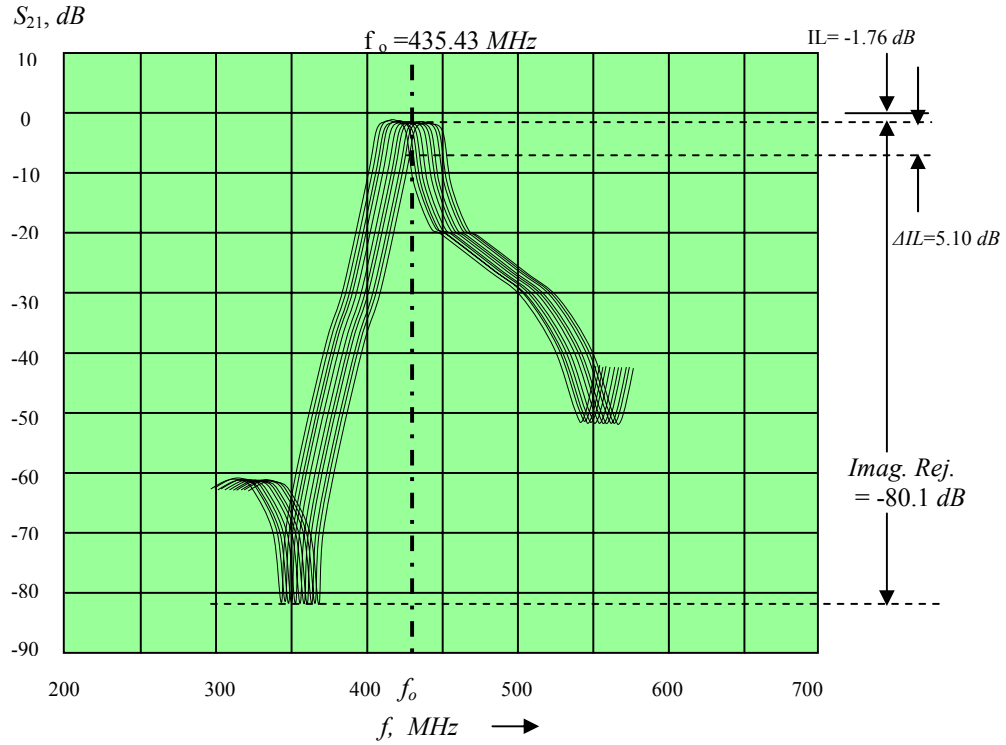
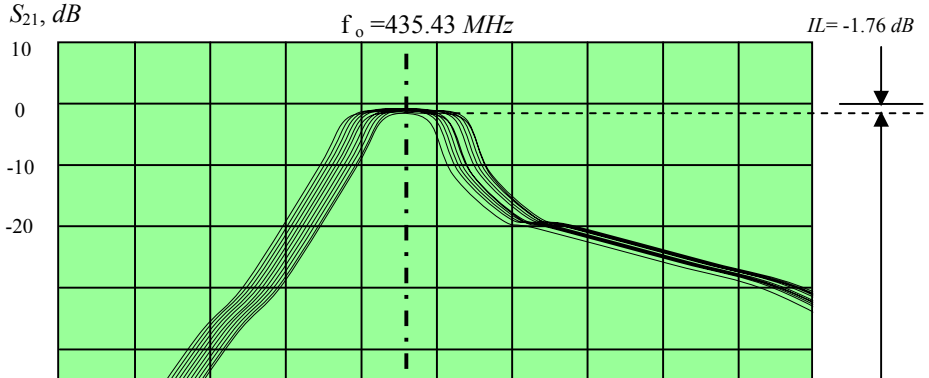


Figure 7.16 Frequency response of tunable filter with tolerance
--- Bandwidth is too narrow

From Figure 7.16 it can be seen that the performance of the product has problem of the bandwidth if the tolerance of parts is considered. At the operating frequency, f_o, the insertion loss, -1.76 *dB*, is just for a few units of the product. Most of units have the insertion loss more than -1.76 *dB*. It could reach up to 5.10 *dB* for some units. Obviously it is not a 6σ design. The yield rate might be pretty low in the production line.

To reach the goal of a 6σ design, all the curves at the operating frequency must have insertion loss of around -1.76 *dB*. Should the bandwidth be increased, the insertion loss around -1.76 *dB* can be reached for all the curves, which is shown in Figure 7.17.

However, the increase of bandwidth is just the opposite from the expectation of "to narrow bandwidth of the tunable filter as much as possible". The bandwidth shown in Figure 7.17 is too wide. An appropriate approach is depicted in the Figure 7.18, where at



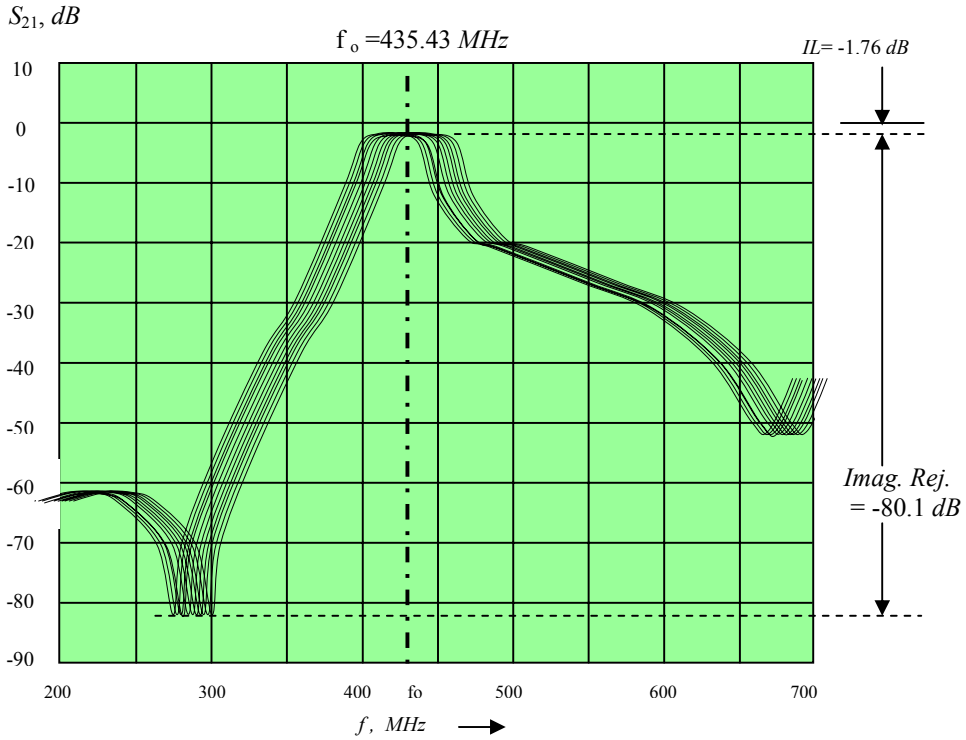


Figure 7.18 Frequency response of tunable filter with tolerance

---Bandwidth is appropriate---

the operating frequency, the insertion loss of -1.76 *dB* is reached for all the curves while the bandwidth is minimum.

From Figure 7.16 to 7.18, it can be concluded that the bandwidth of a tunable filter is determined by the tolerance of the parts. A designer must do the tolerance analysis with other design works to approach the goals simultaneously. The less the tolerance of parts, the narrower the bandwidth of tunable filter can be. On the other hand, the less the tolerance of parts, the higher the cost. A trade-off between the bandwidth and the cost must be taken through the tolerance analysis.

By means of Monte-Carlo analysis, a performance histogram of the circuit block within its specified range and the yield rate can be obtained. Figure 7.19 shows the displayed result after a Monte-Carlo analysis for the insertion loss less than -2.5 *dB*. It can be seen from the histogram that the insertion loss is located between -1.71 and -2.16 *dB*, which is less lossy than the goal, -2.5 *dB*, and therefore its yield rate is 100%.

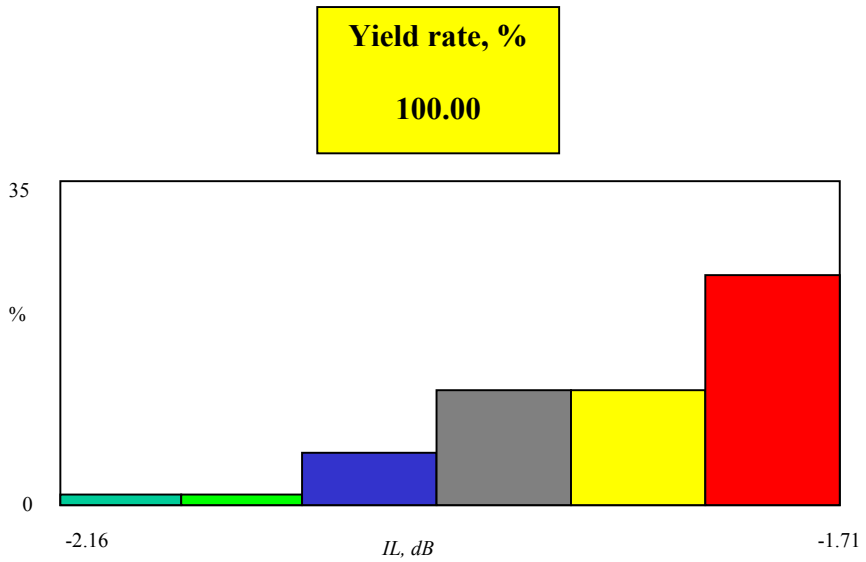


Figure 7.19 Display of insertion loss histogram and yield rate for IL> -2.5 *dB*

Monte-Carlo analysis is one of powerful tools to adjust the tolerance desired for each part. It displays the effect or the sensitivity of the individual part's value on the total performance. Figure 7.20 shows the effect of the 5 individual parts, C_1, C_2, L_1, L_2, and L_3, on an insertion loss which is less lossy than -2.5 dB. 50 iterations are conducted in the analysis so that there are 50 points on the plot. Each point therefore represents one random value of the part in one time of iteration. In the plots for *IL* vs. C_1 and C_2, the tolerance of C_1 and C_2 looks adequate. All the points are homogeneously scattered within the range of insertion loss. On the contrary, in the plots for *IL* vs. L_1, L_2, and L_3, the tolerance of L_1, L_2, and L_3 looks somehow smaller than what is needed because all the points are crowded in the middle

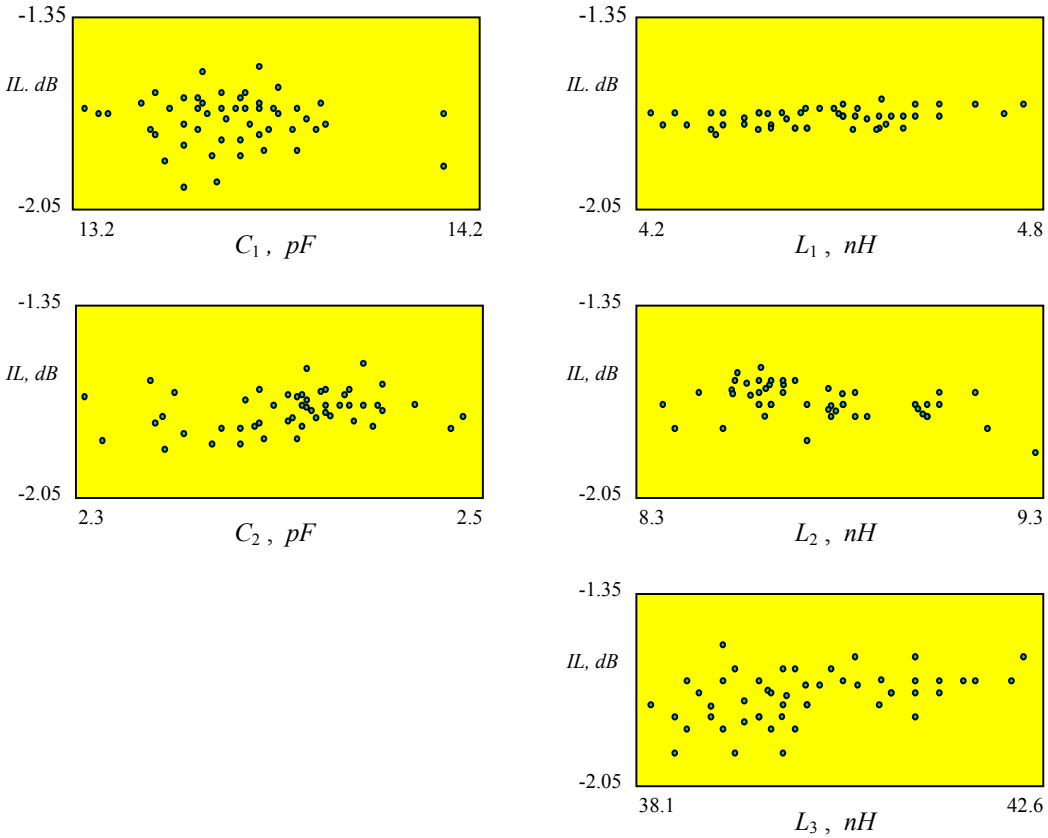


Figure 7.20 The effect of individual part's value on the insertion loss

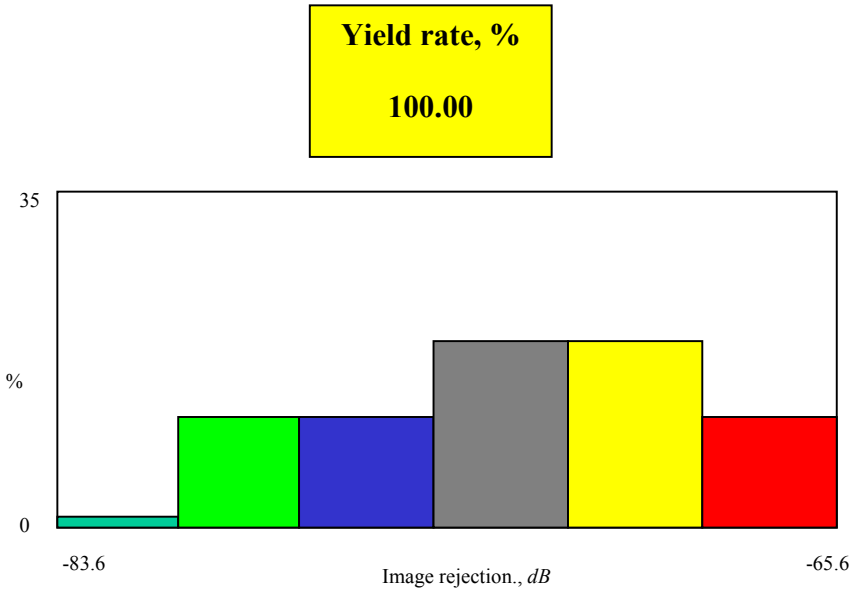


Figure 7.21 Display of image rejection histogram and yield rate for IMRJ< -60 *dB*

area within the *IL* range. Their tolerance, especially L_1's, could be relaxed to a somehow higher value so as to cut the cost down in the discrete design or reduce the die area down in the IC design.

Similar to the analysis for the insertion loss, the analysis for the image rejection is shown in Figure 7.21and 7.22. Figure 7.21 shows the displayed result after a Monte-Carlo analysis for the image rejection which is lower than -60 dB. It can be seen from the histogram that the image rejection is located between -65.6 and -83.6 dB, which is much lower than the goal, -60 dB, and therefore its yield rate is 100%. Figure 7.22 shows the effect of the 5 individual parts, C_1, C_2, L_1, L_2, and L_3, on the image rejection which is lower than -60 dB. 50 iterations are conducted in the analysis so that there are 50 points on the plot.

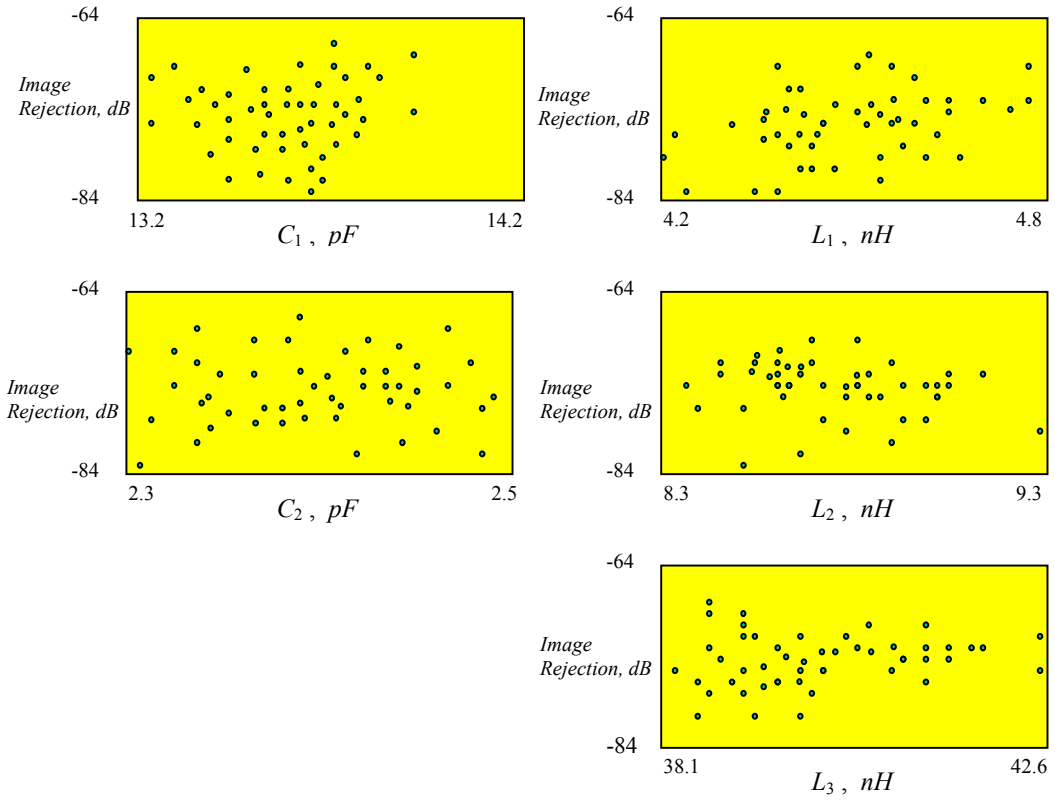


Figure 7.22 The effect of individual part's value on the image rejection

Each point therefore represents one random value of the part in one time of iteration. In all the plots for image rejection vs. C_1, C_2, L_1, L_2, and L_3, the tolerance of all the parts looks adequate because in all the plots, the points are homogeneously scattered within the range of image rejection. The input tolerances for every part are appropriate and have no room to relax for cost reduction.

7.5 Appendix: Table of the Normal Distribution

$$f(z) = \int_0^z \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx = \frac{1}{2} erf\left(\frac{z}{\sqrt{2}}\right)$$

Z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3016	.3133
0.9	.3135	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4778	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4925	.4929	.4931	.4932	4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

References

[1] William H. Beyer, "Standard Mathematical Tables," (Book), CRC Press, 24th Edition, 1976.

[2] Taguchi, G., "Experimental Designs," 3rd Edition, Vol. 1, (In Japanese), Maruzen Publishing Company, Tokyo, Japan, 1976.

[3] Taguchi, G., "Experimental Designs," 3rd Edition, Vol. 2, (In Japanese), Maruzen Publishing Company, Tokyo, Japan, 1977.

[4] Taguchi, G., "Off-Line and On-Line Quality Control System," Proceedings of International Conference on Quality Control, Tokyo, Japan, 1978.

[5] Juran, J. M., "Quality Control Handbook," (Book), 3rd Edition, McGraw-Hill, New York, 1979.

[6] Brayton, R. K., Director, S. W., and Hachtel, G. D., "Yield Maximization and Worst-Case Design with Arbitrary Statistical Distributions," Ieee Transactions on Circuits and Systems, CAS-27, pp. 756-764, 1980.

[7] Donald R. J. White, "Electrical Filters," (Book), Don White Consultants, Inc., 1980.

[8] Phadke, M. S., Kackar, R. N., Speeney, D. V., and Grieco, M. J.,"Off-Line Quality Control for Integrated Circuit Fabrication Using Experimental Design," The Bell System Technical Journal, 62, pp. 1273-1309, 1983.

[9] Raghu N. Kackar, "Off-Line Quality Control, Parameter Design, and the Taguchi Method," Journal of Quality Technology, Vol. 17, No. 4, pp. 176-188, October, 1985.

[10] Victor E. Kane, "Process Capability Indices," Journal of Quality Technology, Vol. 18, No. 1, pp. 41-52, January, 1986.

[11] Richard C. Li, "Tunable Filter Having Capacitively Coupled Tuning Elements," U.S. Patent 5,392,011, Motorola Inc., 1992.

Index

Contents

Chapter 8 Prospect of RFIC Design

8.1 History of RFIC Development

The sophisticated characters of semiconductors were discovered in the 1940's and thus vacuum tubes were rapidly replaced by transistors in the circuit design. And then, since the 1950's, the IC (Integrated Circuit) technology was developed for both analog and digital circuits at low frequencies to high frequencies even to RF (Radio Frequency) range. However, the RF circuits have still been implementing with discrete parts up to 1980's.

At high or radio frequencies, the conceivable electromagnetic field radiates from each part. Therefore, every part in the circuit, including each runner, looks like a small antenna. All that radiation could mingle to form serious interference which would cause disorder in the circuit and put out of work. Thus, various shielding technologies have been developed to isolate a block from other blocks. In the electronic industry history isolation was and is still one of key segments in the RF circuit design. However, the cost for isolation is very high due to the expensive shielding materials and the size of circuit block gets huge due to the shielding box or some other isolation equipment. The implementation of RF circuits becomes a bottleneck in an electronic product when RF blocks are needed.

In order to reduce the cost, scientists in the laboratories and engineers in the R&D departments initialized the RFIC (Radio Frequency Integrated Circuit) development since 1980's, that is, to put RF blocks into the integrated circuits, just like they did for both of analog and digital circuits operating at a low frequency. It was a very brave action in the history of the electronic industry.

In early stages, most experiments were concentrated in the isolation between the devices. The trench scheme, and then the guard ring technology were developed one by one to enhance the isolation. The experimental results were very encouraging. Eventually the isolation research shifted from the devices to the blocks. In the 1990's, the isolation between blocks had been worked out in many laboratories and R&D departments in

Table 8.1 Some of companies developed the 1st RFIC chip in 1995-1996 in the United States

Company	**Philips**	**Motorola**	**Tek/Maxim**	**Harris**	**AD/IBM**
Technology	QB1C1 GST-2 C-PI	MOSIC-5 BiCmos85	QUICKCHIP-8	UHF-1	SiGe
Ft(NPN)	13 GHz	12/12 GHz	12/27/12 Ghz	8.6 GHz	50 GHz
Ft(PNP)	900MHz	700/50M2	8.5G/50M/8.	5GHz	4GHz
Status	Pro	Pro	Pro/Dev/Pro	Pro	Dev
Cycle time1	90 days	50 days	30-70 days	90 days	90 days
Risk	L/M-H/L-M	M-H/M-HM	M-H	M-H	M-H
$/wafer	TBD	$1.2k/2k	$10k	TBD	$2k
Design sys.	-	Cadence	QUICKIC	FASTRAC	SPICE
RF IC	SA601/620	MC13142		3600	

many universities, institutes, and companies. Table 8.1 lists some of them in the United States.

As the 1st RFIC chip, Philips put an LNA (Low Noise Amplifier) together with a mixer. The performance of both of the LNA and the mixer were pretty good and the isolation between these two blocks reached 40 dB! Table 8.2 shows their performance.

Table 8.2 The 1st RFIC chip developed by Philips in 1996

Philips' Chip SA601@ 881 MHz		
Block	**LNA**	**Mixer**
Power supply	3 V	3 V
Current drain	3.0 mA	4.4 mA
Gain	11.5 dB	6 dB
Noise Figure	1.6 dB	9.5 dB
IIP3	-2 dBm	-2dBm
Isolation	40 dB	

(*From LNA output to Mixer RF input.)*

An isolation of 40 dB was an unprecedented record though it is still far from the ideal goal. The perfect isolation, of course, is infinitive dB. At the same time, some other companies, such as Motorola and others, were working for the same objective. The isolation, more or less than 40 dB, was obtained. At that period, many processes have been developed. Table 8.3 shows some of them.

Table 8.3 The various processing in the development of RFIC from1998 to 2000

Processing	**F_T**	**Reliability**	**Cost**
SiGe BiCMOS	50-75 GHz	OK	M+
Si CMOS	15-25 GHz	Good	L
Si BiCMOS	15-35 GHz	Good	M
GaAs HEMT	50-90 GHz	OK-	H

In 1998, IBM's laboratory developed the new Silicon-Germanium processing (SiGe BiCMOS) . Their RFIC chips were developed as shown in Table 8.4.

Table 8.4 RFIC chip set developed in IBM via SiGe BiCMOS in 1998

IBM	**LNA SGRF0100**	**VCO SGRF4111**	**LNA + IR Mixer*) SGRF2113**
Vcc	3V	3V	3V
Icc	5.5 mA	10.0 mA	15.0 mA
Frequency	1.9 GHz	880/1850 MHz	900/1900 MHz
Gain	G =12.5 dB		G = 22 dB
Noise Figure	1.2 dB		3 dB
SBN/Phase noise		-92 dBc/Hz @ 20 kHz offset	
IIP_3	13 dBm		-12 dBm
k		120 MHz/V	
P_{out}		-10 dBm	
Isolation			20 dB

*) *P_{LO} = -10 dBm, IF = 400 MHz, Load_IF = 600 Ω.*

The outstanding performance was the low noise figure in the LNA and the low sideband noise or phase noise in the VCO. The low noise characterization and other good features are very attractive to the RF designers up to now.

Since then, the development of RFIC chips has accelerated. Many good products have been announced or are available in the market. Table 8.5 lists some of them since the year 2000. Table 8.6 lists some RFIC chips developed for the WLAN IEEE 802.11a system.

Table 8.5 Some of RFIC chips developed since 2000

Company	**Infineon (2000)**		**RF Micro Devices (March 2001)**	**Silicon Labs. in Austin (March 2001)**	**Qualcom (February 2001)**
	LNA	**PA**	**2 LNAs / 3 Mixers RF2489**	**RF IC chip, CMOS**	**RF T5200 , RFR52000**
Vcc	*3V*	*2.8V*			
NF	*1.1 dB*		*1.1 to 1.3 dB*		
Freq.	*1.9 GHz*	*900 MHz*			
Gain	*28 dB*		*15 to 18 dB*		
P_{out}		*3.2w*			
Eff.		*54%*			
IIP_3			*-4 to 12 dBm*		
Cost			*$2.31 for 100k pieces*		
Mainly for			*WCDMA/ GPS*	*GSM /GPRS*	*WCDMA*

Table 8.6 Some of RFIC chips for WLAN IEEE 802.11a system

Company	**Intersil**	**Resonext**	**Bermai**	**Atheros**	
Freq. Band	5.15-5.25	5.15-5.25	5.15-5.25	5.15-5.25	GHz
	5.25-5.35	5.25-5.35	5.25-5.35	5.25-5.35	GHz
		5.725-5.82			GHz
Rx	Dual-Conv.	Zero IF		Dual-Conv.	
Tx	PA for indoor	18 dBm PA	22 dBm PA		
	External PA	External PA	W/O Ext. PA	W/O Ext. PA	
Package	MLF	MLF		LLPC	
Processing	CMOS	0.18μmCMOS SiGeBiCMOS	0.18μmCMOS	0.25μmCMOS	
# of chips	3 +PA+SAW	2	1	2	

8.2 Isolation between Blocks in an RFIC

8.2.1 Definition and Measurement of Isolation

A perfect isolation between blocks implies that there is no cross-talk or interferences between blocks. It can be measured by the measurement of the signal attenuation from point A to B as shown in Figure 8.1.

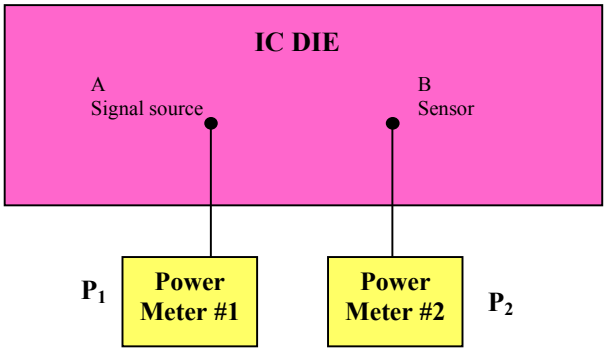


Figure 8.1 Definition of isolation between A, the signal source, and B, the sensor

Assuming that the power of the signal source at point A is P_1 measured by the power meter #1. The power of sensor at point B is P_2 measured by the power meter #2. Both of them are expressed by dB. The isolation between A and B is defined as

$$Isolation = 10\log\frac{P_1}{P_2},\ \text{dB} \tag{8.1}$$

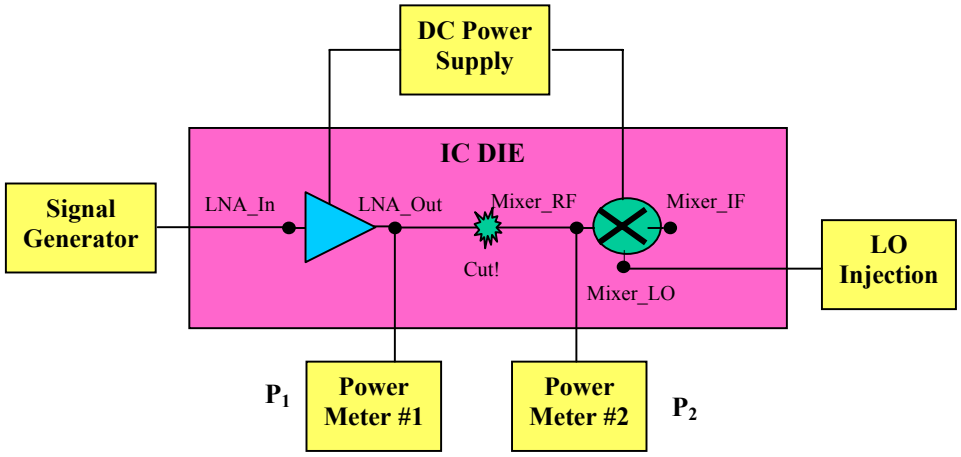
A perfect isolation is when

$$P_2 = 0\ , \tag{8.2}$$

then,

$$Isolation \to \infty\ . \tag{8.3}$$

Figure 8.2 shows the measurement of the isolation between the RF output of the LNA and the RF input of the mixer. These two points, the RF output of the LNA and the RF input at the mixer, are usually connected directly or by a matching network. In order to measure its isolation, the connection is intentionally cut and the powers, P_1 and P_2, are measured when the LNA and the mixer are in a normal operating state. That is, the DC power supply must be provided to both LNA and mixer, the RF signal from a generator must be fed to the input of the LNA, and the appropriate LO injection must be connected to the LO portion of the mixer.



$P_1 - P_2$ = Isolation, dB, from the LNA_out to the Mixer_RF , when the LNA and the mixer are in a normal operation state.

Figure 8.2 Relationship between source and load

8.2.2 Isolation Technology

The bottleneck of an RFIC technology is the isolation between the RF blocks. After a number of hard works for many years, it is found that the main ways, by which the strong interference or cross-talk between RF blocks results the poor isolation, they are:

- The common substrate of IC die;
- The electromagnetic field from inside and outside the block;
- The DC power supply line;
- The related pads of blocks;
- The bonding wires.

In the following sections let's introduce the remarkable works in the improvement of the isolation between RF blocks.

1) Guard ring or trench on substrate

Experiments indicate the main contribution of cross-talk between RF blocks comes from the common substrate of the IC die. Most of the electromagnetic field is condensed in the substrate but not in the air because the electricity constant ε_r in the substrate is much higher than in the air.

In the early stages of RFIC design, the intuitional solution of the cross-talk due to the substrate is to cut the substrate away as much as possible. This is the so-called "trenching

of substrate" technology. Figure 8.3 shows the rectangular trench frame around every block.

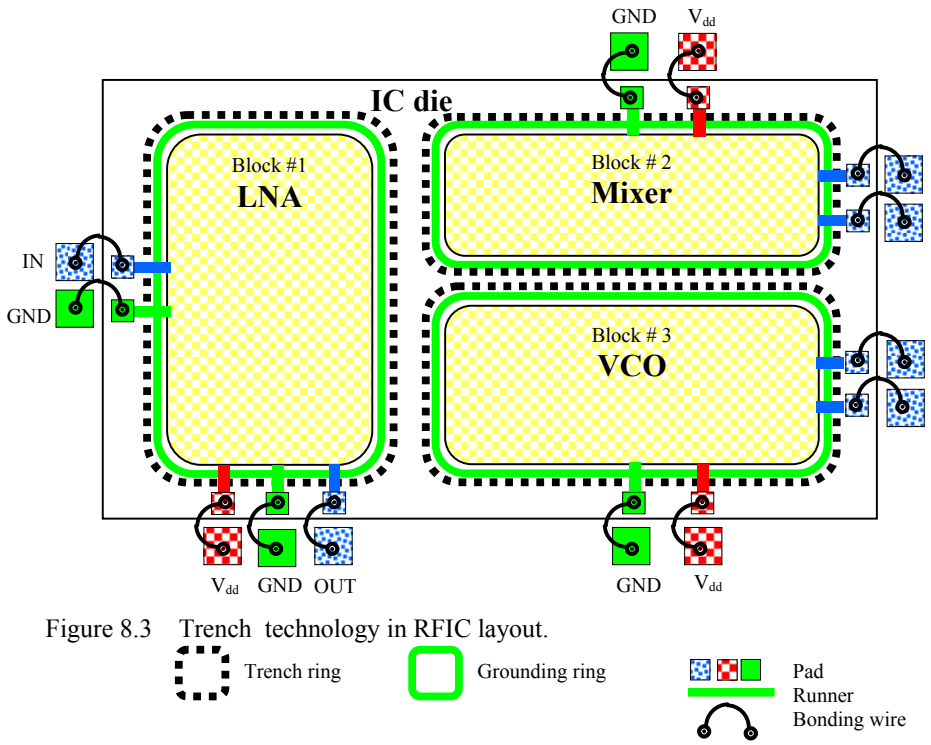


Figure 8.3 Trench technology in RFIC layout.

Its width, depth, plus the spacing from the circuit parts have been experimentally conducted and a number of data have been piled up over the years. It pushed the isolation technology forward and opened an encouraging prospect. However, its process is somehow complicated and it causes mechanic weakness if the width and depth of the trench is too wide or too deep. Later on, people replaced the trench technology by the fabrication of a guard ring. Either P+ guard ring or N-well guard ring has been tested and very positive results have been obtained.

The experiments have been conducted in 3 aspects:

a) Effects of different types of guard rings on cross-talk;
b) Effects of the guard ring parameters, spacing S, width W, and distance between source and sensor D, on cross-talk;
c) Effects of multiple guard rings on cross-talk.

The signal from its port 1 serves as a cross-talk source.

Figure 8.4 shows that the cross-talk is sensed and detected by a network analyzer. The DUT is a simple IC die mounted on a small PCB. The simple IC die, as shown in the magnified portion in Figure 8.4, consists of only two parts on the substrate: One is an n-MOSFET and another one is a P+ buried pad. The gate of n-MOSFET device is provided a signal from the port 1 of the network analyzer and functions as a cross-talk source,

while its source and drain are connected together and functions as a ground termination of input and output. The P+ buried pad functions as a cross-talk sensor. It has about the same size as that of n-MOSFET gate and is located about 100 μm away from the n-MOSEFET. The sensed cross-talk signal is connected to the port 2 of the network analyzer. The parameter S_{21} represents the isolation from the cross-talk source to the cross-talk sensor, that is,

$$\text{Isolation} = -S_{21}\ \text{dB}. \qquad (8.4)$$

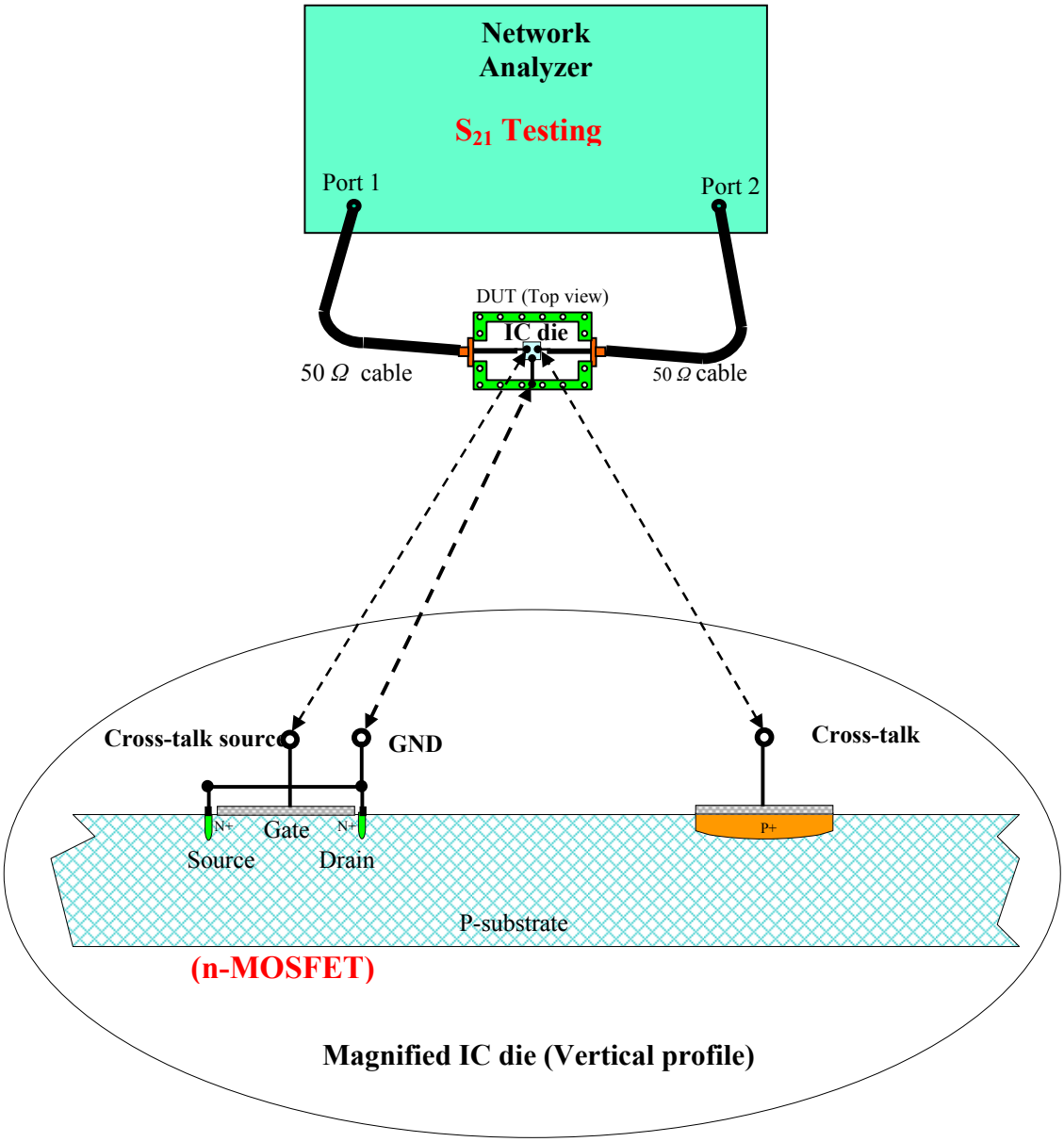


Figure 8.4 Cross-talk between cross-talk source and sensor

N+ P+ P substrate Contact

The experiment (a) as shown in Figure 8.4 is taken as a reference, in which the cross-talk is "original" without any means to improve isolation. The experiments (b) and (c) are designed for isolation improvement or cross-talk reduction and are shown in Figure 8.5 and 8.6 respectively. Figure 8.5 shows that the cross-talk source on the IC die is encircled by a P+ guard ring. Figure 8.6 shows that the cross-talk source on the IC die is encircled not only by a P+ guard ring but also by a N-well.

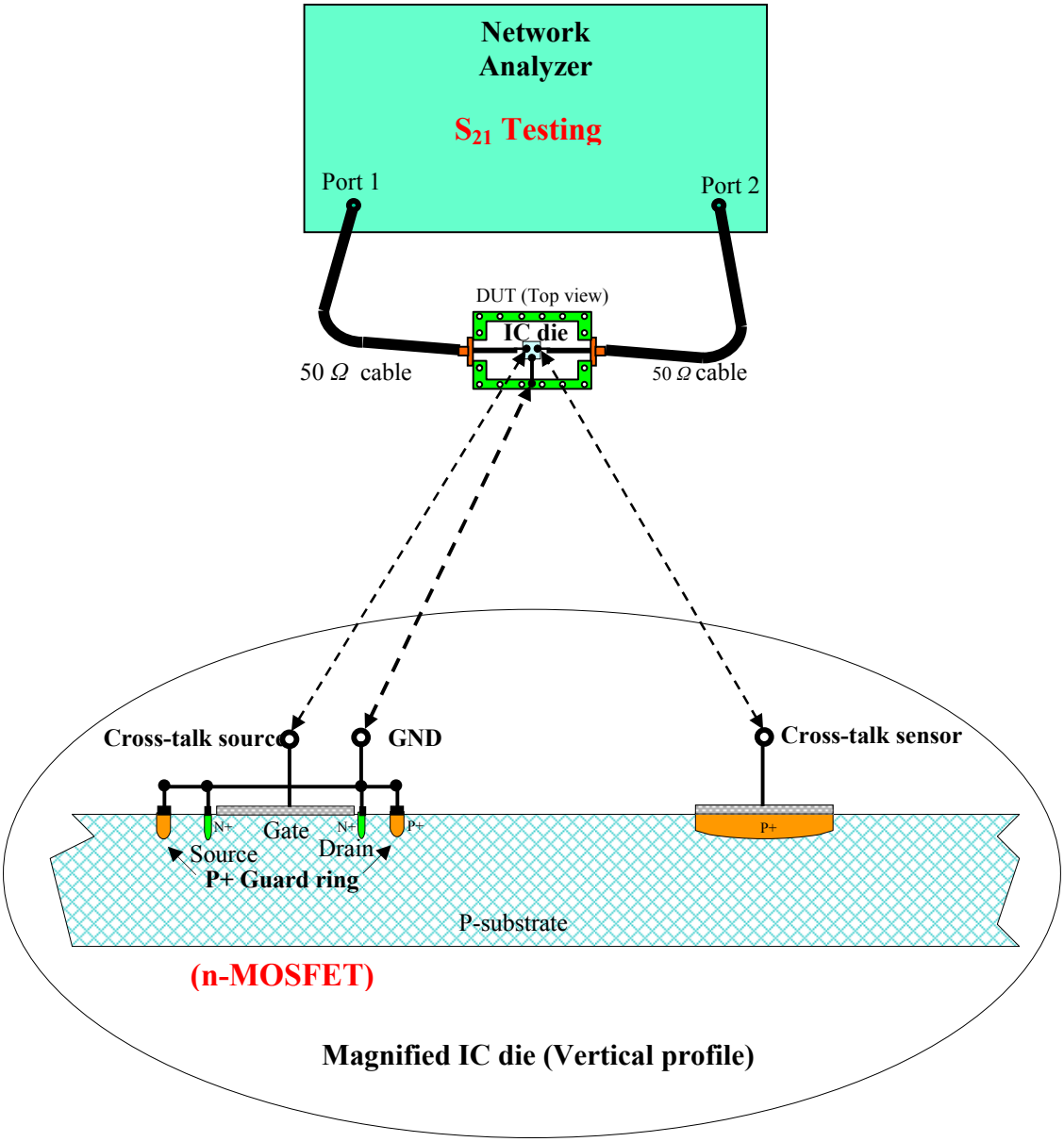
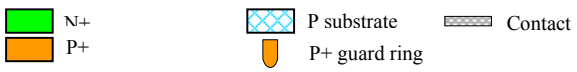


Figure 8.5 Effect of P+ guard ring on cross-talk



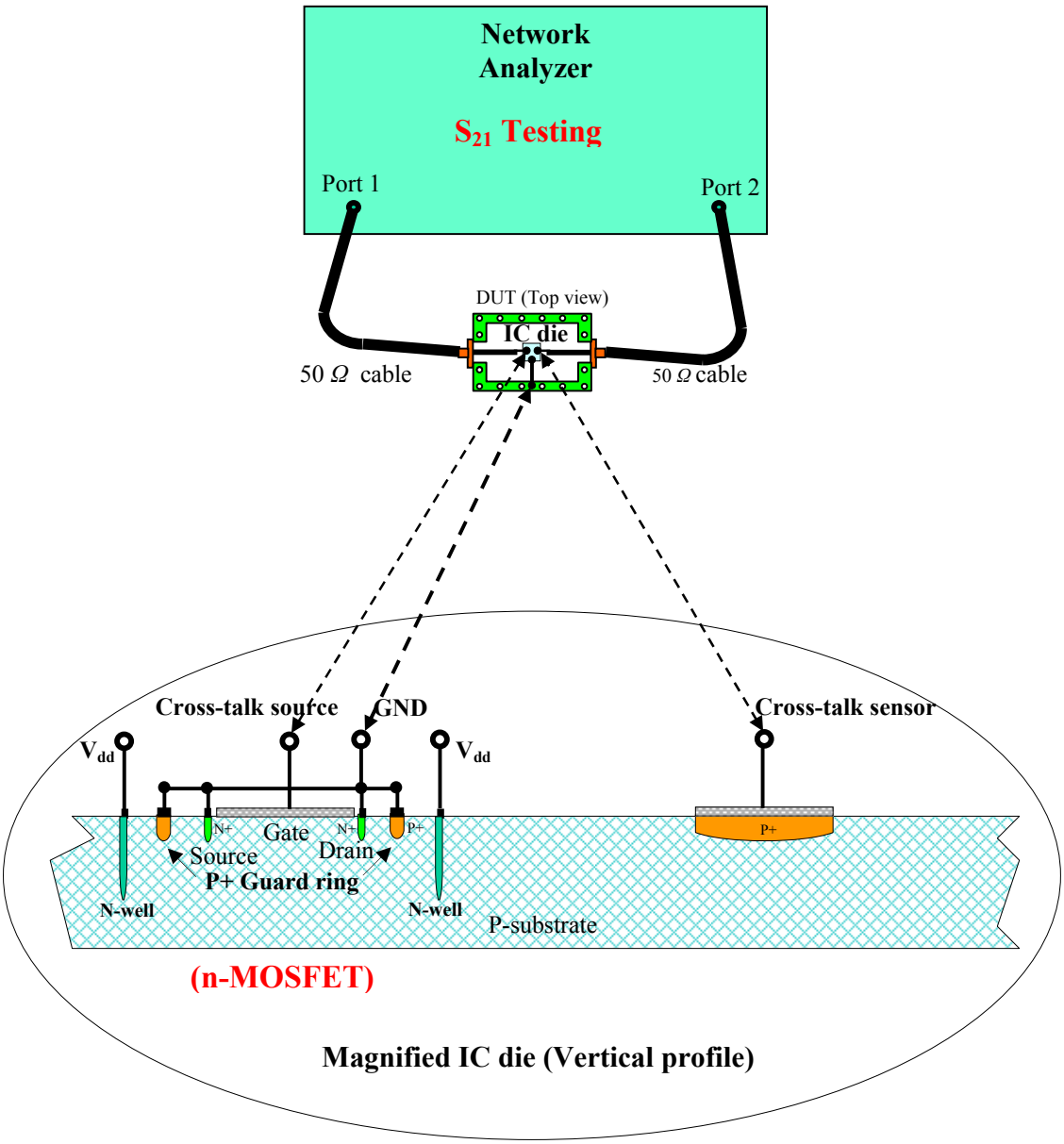
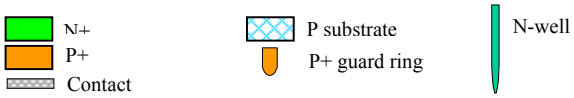


Figure 8.6 Effect of P+ guard ring and N-well on cross-talk



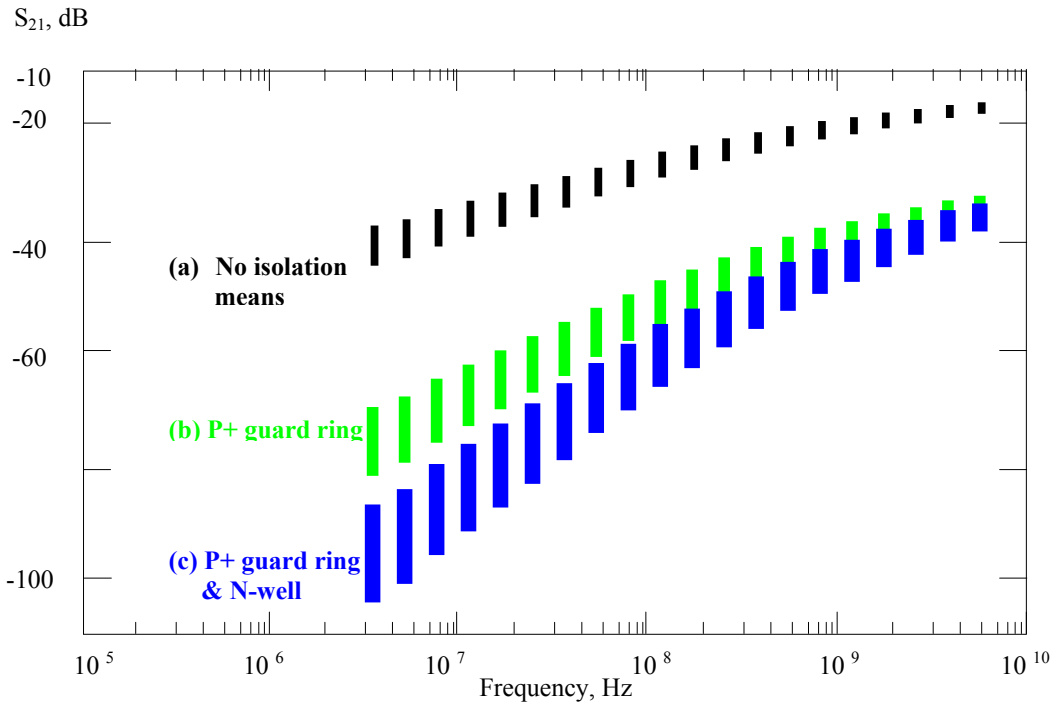
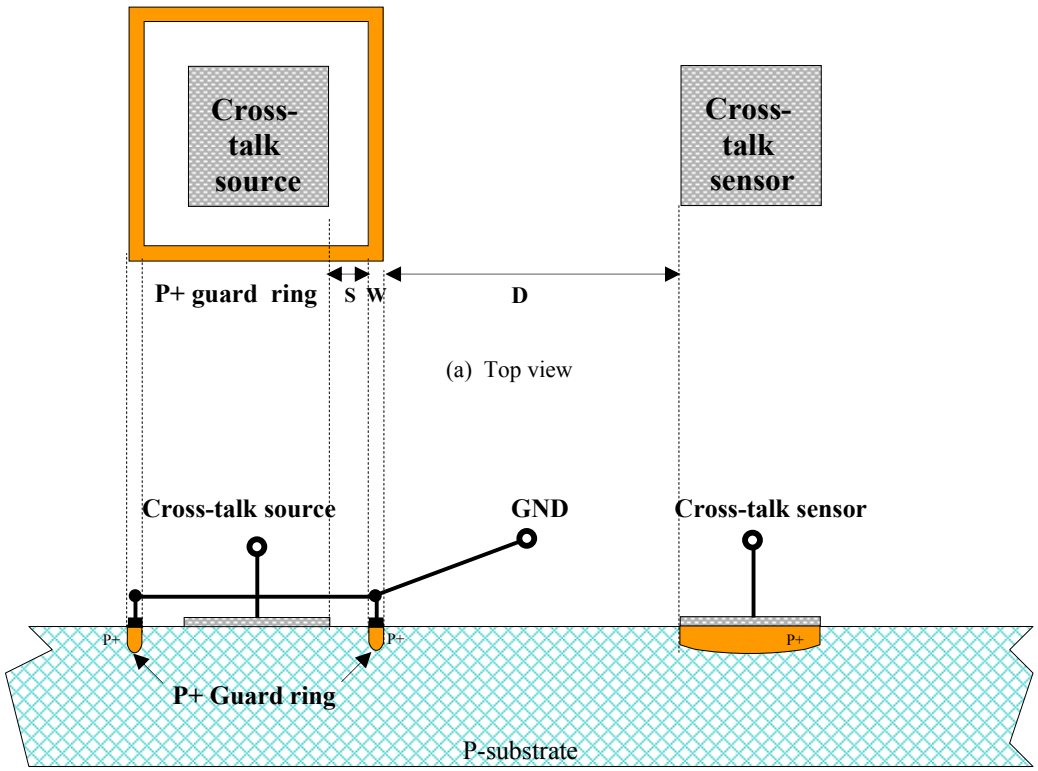


Figure 8.7 Measured isolation, S_{21}, vs frequency for different isolation schemes

The tested results are shown in Figure 8.7. In general, isolation becomes worse as the frequency is increased in all experiments. Without any isolation protection as in experiment (a), the isolation is more than 30 dB at 10 MHz but only about 20 dB at 1GHz. The results from experiment (b) are quite encouraging and exciting. At 10 MHz, the isolation is increased up to 70 dB and at 1 GHz is increased up to 40 dB approximately. The experiment (c) is an attempt to furthermore improve the isolation with an additional N-well encircling around the P+ guard ring. The outcome is positive as expected.

The next objective is to study the effect of the various parameters of the guard ring on the isolation. As shown in Figure 8.8, the main parameters of the guard ring include the width of the guard ring, W, the spacing between the guard ring and the cross-talk source, S, and the distance between cross-talk source and sensor, D.

It is found that the effect of the spacing between the guard ring and the cross-talk source, S, on the isolation is not too sensitive if the operating frequency is about or over 1 GHz, while it becomes different as the operating frequency is lower than 1 GHz. The isolation is better as the spacing is narrowed. The narrow spacing is beneficial to minimize the die size also.



(b) Vertical profile

Figure 8.8 Effect of guard ring parameters, S, W, D, on cross-talk

It is found that the isolation becomes better if the width of the guard ring is wider. By testing examples, the isolation could be increased 10 to 20 dB if the width of the guard ring is increased from 4 μm to 64 μm. The disadvantage of a wider guard ring is that the die size might be increased up to an impractical size, which might be too expensive. The designer should take a trade-off between the necessity of the isolation level and the cost limitation. For an RF block with low power operation, a 10 μm width of guard ring might be enough to reach the specified isolation level. On the contrary, for the PA design, a width of guard ring more than 50 μm might be desired so as to reach and over the specified isolation level.

It is found that the isolation becomes better if the distance from guard ring to the sensor is longer. By testing examples, the isolation could be increased by 10 to 20 dB if the distance from guard ring to the sensor is increased from 50 μm to 500 μm. In reality it is very hard to increase the distance over, say, 50 μm, since the die size is limited. The designer should take a trade-off between the necessity of the isolation level and the cost limitation. For the RF block with low power operation, a distance shorter than 50 μm

from guard ring to the sensor might be enough to reach the specified isolation level. On the contrary, for the PA design, a longer distance, say, over 100 μm, from guard ring to the sensor might be not enough to reach the specified isolation level.

The variation of the isolation due to the change of the width of the guard ring W, the spacing between the guard ring and the cross-talk source S, and the distance between cross-talk source and sensor D, has been reflected in Figure 8.7, where each value of S_{21} is varied within a range and displayed by many short vertical bars.

Finally, we will study the effect of multiple guard rings on the isolation. Figure 8.9 shows 3 guerd rings circling around the cross-talk source. Better isolation is usually obtained as the number of guard rings is increased. The price is the larger die size as the number of guard rings is increased.

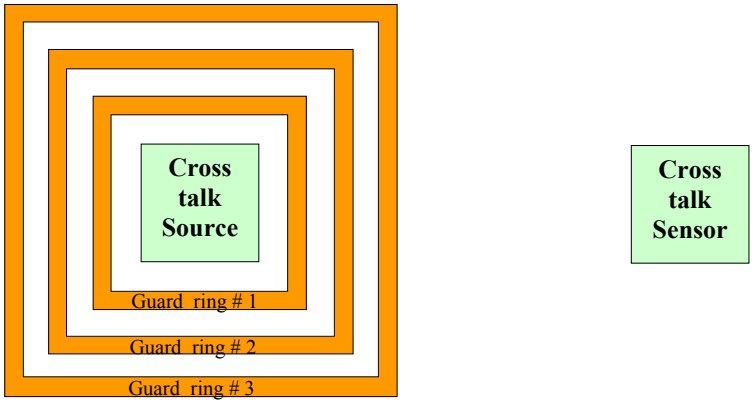


Figure 8.9 Effect of multiple guard rings on cross-talk.

So far we have discussed how to minimize the cross-talk due to the substrate.

At present many designers are applying one P+ guard ring and one N-well in their RFIC layout. The geometrical parameters of the guard ring and the N-well are selected on the basis of the trade-off between the cost and the performance of isolation. Figure 8.10 shows a set of recommended values of these parameters.

The experiments indicate that the better performance of isolation is when the N-well guard ring is connected to the DC power supply, Vdd, and the P+ guard ring is connected to the ground as shown in Figure 8.10.

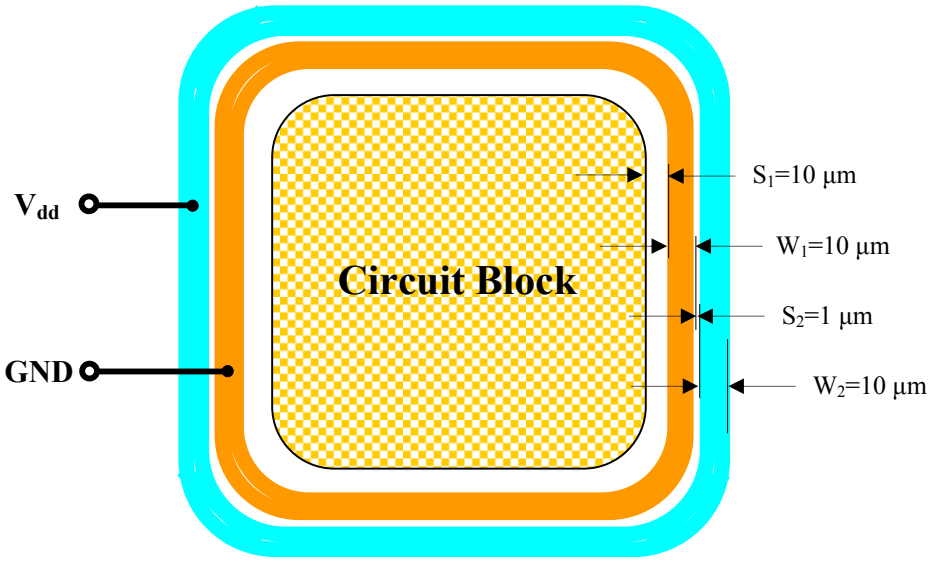


Figure 8.10 Recommended width , spacing and other parameters for of P+ and N-well guard ring.

P+ guard ring N-well guard ring

2) Grounding ring separately

The guard ring is for the reduction of cross-talk throughout the substrate, which is the main portion of the interference because the electric permittivity of the substrate usually is much higher than that of the free space.

The second source of cross-talk is the interference of the electromagnetic field around the blocks, which is in the space above the IC die. A grounded ring circling the block as shown in Figure 8.11 and 8.12 is quite helpful to reduce the "cross-talk" of the electromagnetic field between blocks in the air. The electric lines radiated by a block would be bent down to the grounded ring circling the block so that the electromagnetic field produced by the block would be confined within its local block area, and would not disturb its neighboring blocks. Therefore, a grounded ring circling a block is another one of important means to reduce the cross-talk between the blocks. It is a scheme of so-called "separate grounding for each block".

It should be noted that in both of Figure 8.11 and 8.12, the "separate grounding for each block" is implemented and all of the individual grounded ring must be connected to the grounding terminal of the common DC power supply. Figure 8.11 shows that the connections between the individual grounding rings and the grounding terminal of the

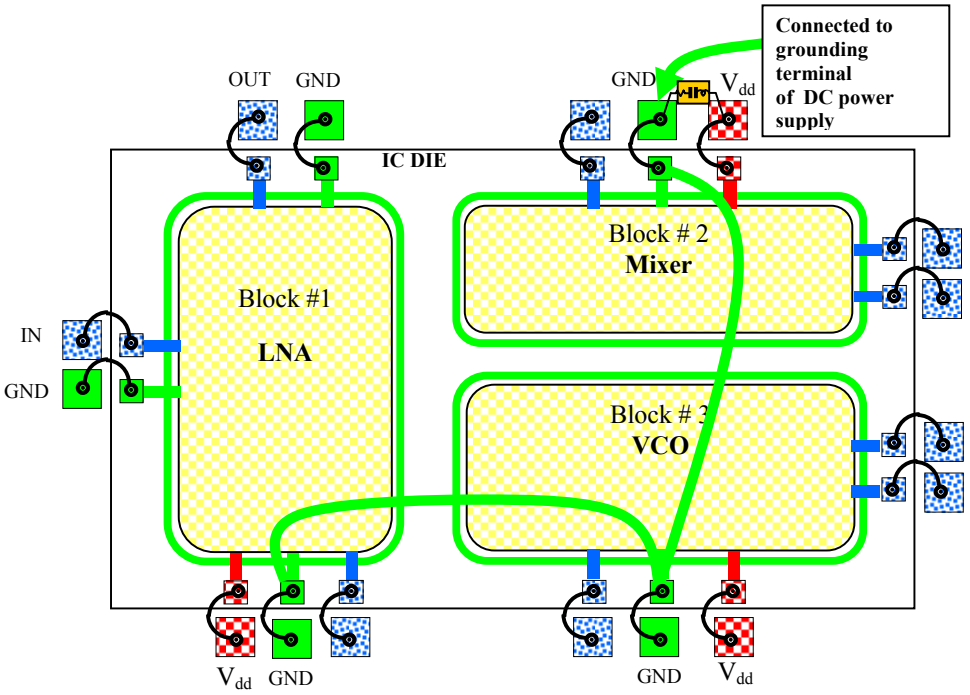


Figure 8.11 Separate grounding for each individual block but incorrect grounding connection to DC power supply.

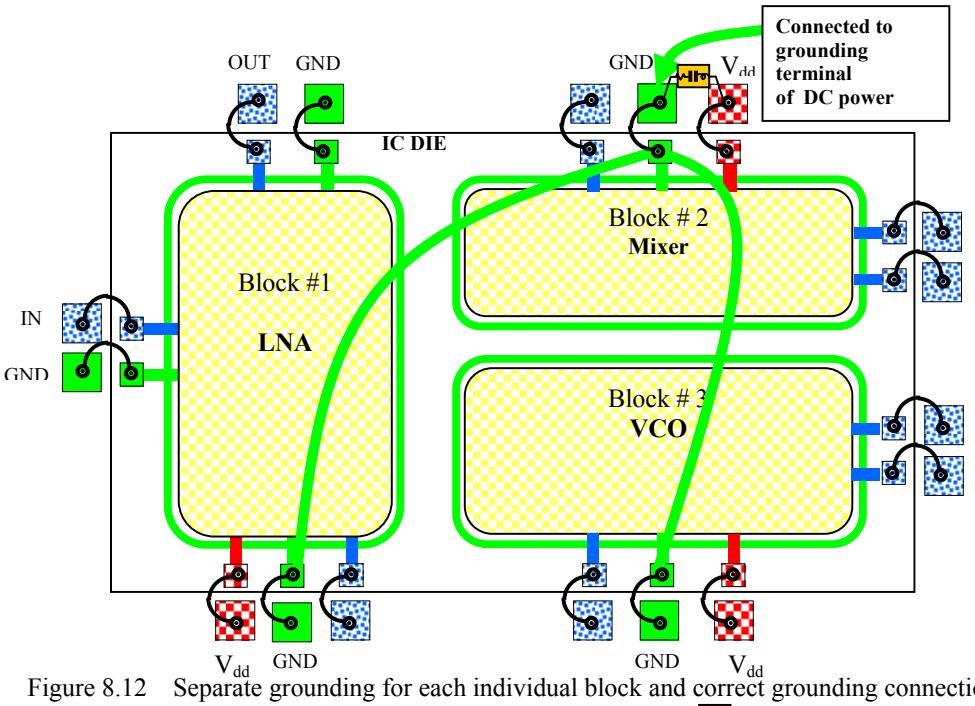


Figure 8.12 Separate grounding for each individual block and correct grounding connection to DC power supply.

common DC power supply are in series while Figure 8.12 shows that the connections between the individual grounding rings and the grounding terminal of the common DC power supply are in parallel. Figure 8.11 shows an incorrect connecting way because, as discussed in Chapter 3, the connection in series will bring about the magnetic coupling in the common return current path. The correct connection is shown in Figure 8.12, in which all the grounding pads, one for each block, are connected to the off-chip common grounding pad in parallel, and in addition, there is a "zero" capacitor connected between the off-chip V_{dd} pad and the off-chip grounding pad, where the unique DC power supply is provided outside the IC die.

Theoretically, the cross-talk due to the interference of the electromagnetic field around the block is reduced as the width of this grounded ring is increased. However, the price is the increased die area to be paid. The spacing between the grounded ring and the circuit block is set as narrow as possible, or as close together as possible. Usually it is anywhere from 5 to 50 μm.

As a matter of fact, instead of specially constructing such a grounded ring, the P+ guard ring mentioned above is usually applied for such a purpose. It must be therefore connected to the individual block's ground.

3) DC power supply separately

The third source of cross-talk between RF blocks is the magnetic coupling through the forward currents from DC power supply if a common DC power supply line is applied to all the individual block. Therefore, the DC power supply must provide to individual block separately as shown in Figure 8.14.

However, Figure 8.13 shows an incorrect connection of the DC power supply though an individual DC power supply pad is provided to all the blocks separately. In Figure 8.13, DC power supply is connected form off-chip DC power supply to block #2, the mixer, and then from block #2 to block #1, finally from block #1 to block #3 in series within the IC die. The runner of the DC power supply within IC die brings about the magnetic coupling through its common forward currents and consequently, results in cross-talk between the blocks.

The connection to off-chip DC power supply from the individual block must be in parallel as shown in the Figure 8.14.

The scheme of "DC power supply separately" can be summarized as follows:

- On the IC die, one DC power supply pad must be provided for each block. All the DC power supply pads on the IC die must be connected to off-chip common DC power supply pad in parallel;.
- Outside the IC die, a "zero" capacitor must be connected between the off-chip common DC power supply pad and the off-chip common ground pad.

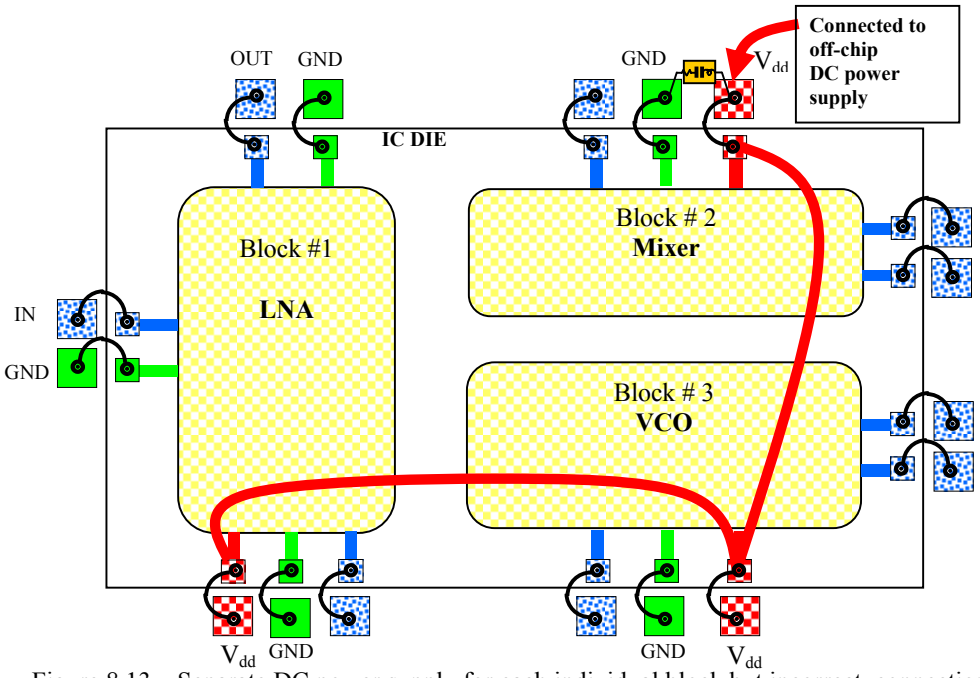


Figure 8.13 Separate DC power supply for each individual block but incorrect connection to off-chip DC power supply

Power supply runner
"Zero" capacitor
Pad
Bond wire

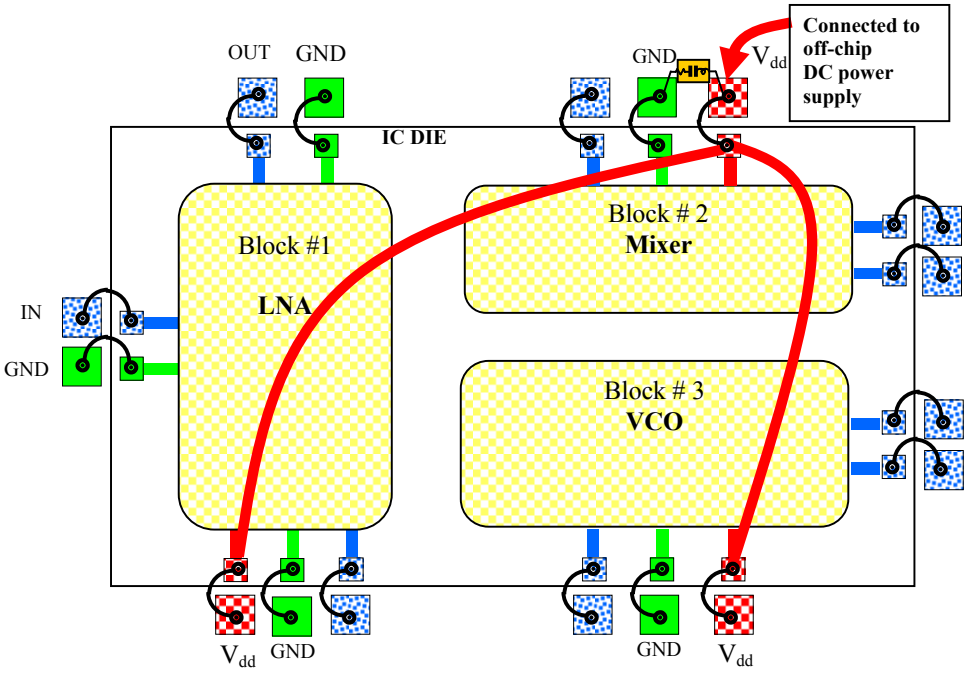


Figure 8.14 Separate DC power supply for each individual block and correct connection to off-chip DC power supply

Power supply runner
"Zero" capacitor
Pad
Bond wire

4) Special RF pad

The bonding pad for an RF signal is different from other pads. The key factor to be considered is its capacitance to the substrate, which attenuates the RF signal to a conceivable amount. Figure 8.15 shows the cross section of an RF bonding pad. The surface of the bonding pad is built with metal 1 and metal 2. These two metals are connected through vias. The area built in metal 1 is reduced down as much as possible so as to lower the capacitance between the metal and the substrate. Under metal 1 is the oxide layer. There is an N well to be created beneath the oxide layer. The N well, N+ buried layer and P substrate forms a P-N junction. Without this N well, the capacitance between the metal and substrate has a high value. The P-N junction inserted between the metal and the substrate greatly reduce the value of the capacitance, especially when the P-N junction is in the reverse-bias state, that is, the N-well is connected to V_{dd} or V_{cc} and the substrate is grounded. The P well and the P buried layer surrounding the N-well is to confine the capacitance contributed only the area beneath the metal 2. The sidewall capacitance under this configuration could be considerably reduced down.

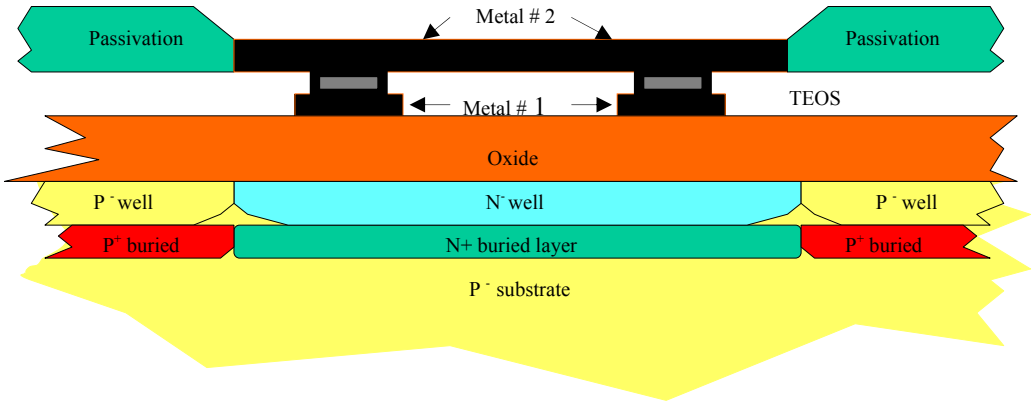


Figure 8.15 RF Bonding pad cross section

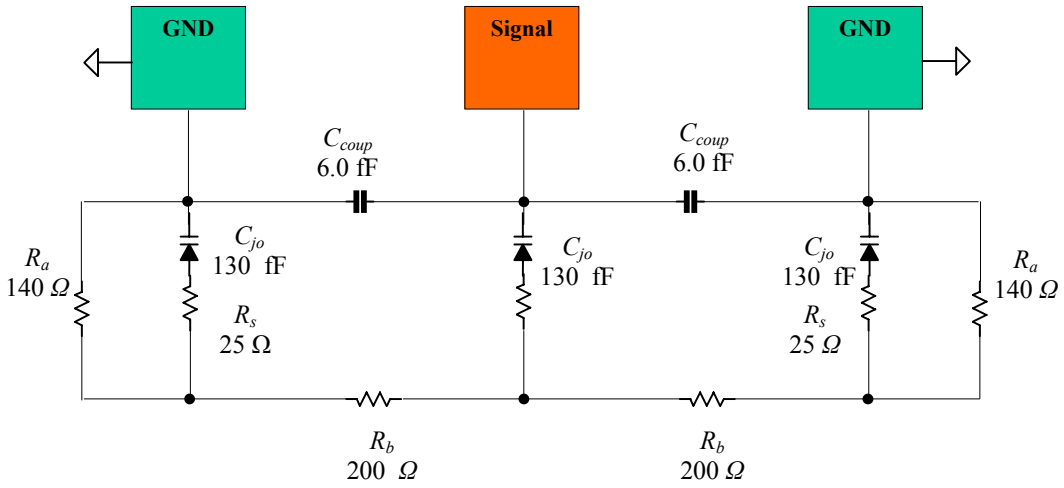


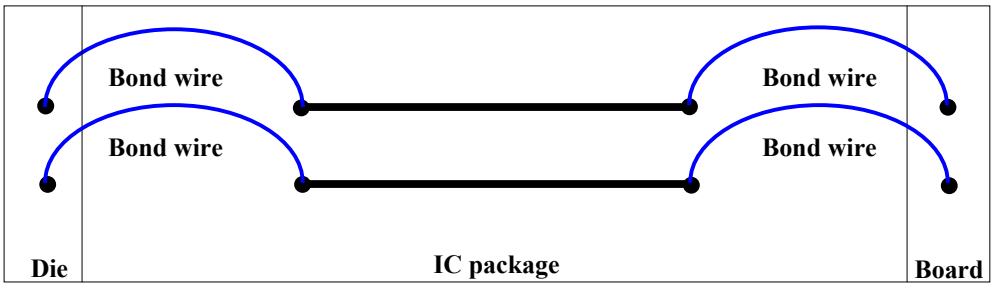
Figure 8.16 An example of the equivalent circuit for open pad test structure

Figure 8.16 shows a typical equivalent circuit of an RF obtained from the test results. The value of the junction capacitor C_{jo} is one tenth pF while the sidewall capacitance is just a couple of fF.

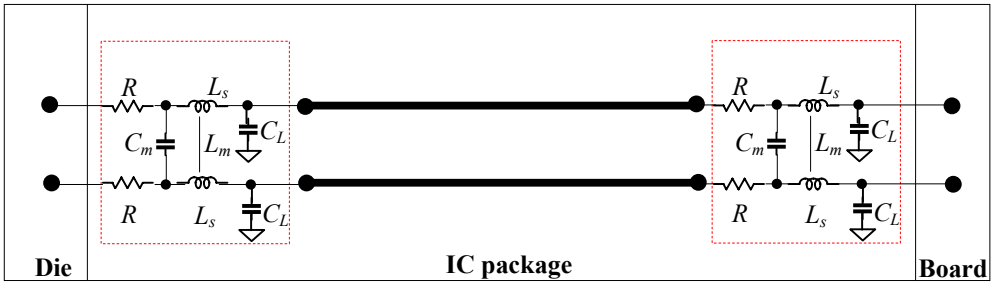
5) Bonding wires

Bonding wires' mutual coupling inductance L_m and mutual coupling capacitance C_m could bring about cross-talk between themselves. The values of L_m and C_m depend on the length, diameter, conductivity, and environment of the bonding wires.

Figure 8.17 shows a typical equivalent circuit of bonding wires, where
R = Resistance of the selected wire;
L_s = Self-inductance of the selected wire;
L_m = Mutual inductance among the selected and adjacent wires;
C_L = Load capacitance from the selected conductor to all other adjacent wires;
C_m = Mutual capacitance among the selected and adjacent wires.



(a) Bond wires from IC die to a Board



(b) Equivalent circuit of bond wires from IC die to a board

Figure 8.17 Equivalent circuit for open pad test structure

Their values depend on

- Package type,
- Package size,
- Die size,
- Frequency,
- Diameter of bond wire,
- Number of bonding wires at each pad.

The values of their parameters are in the range of

$0.1\ \Omega < R < 1\ \Omega$,
$1\ nH < L_s < 2\ nH$,
$0.1\ nH < L_m < 1\ nH$,
$0.1\ pF < C_L < 0.5\ pF$,
$0.01\ pF < C_m < 0.2\ pF$.

It is easy to understand that it is better to shorten the bonding wire as much as possible. The problem is that the length of the bonding wire very often like a wild horse, and very hard to control.

In order to reduce the resistance of the bonding wire, sometimes multiple of bonding wires are applied to a bonding pad. In the practical production line, the maximum wires' number is only two. It is found that the resistance of bonding wire is reduced about 50% if the single bonding is replaced by the double bonding, while the other parameters are not considerably changed.

8.3 Low Q Value of Spiral Inductor

In the RFIC circuit design, there are 4 basic parts: a device (MOSFET, bipolar transistor and so on.), capacitors, inductors, and resistors. Among these 4 dispensable parts, the RFIC engineers have quite a headache about the inductor.

The inductor in an IC chip has typically a spiral configuration as shown in Figure 8.18, and its equivalent is shown in Figure 8.19.

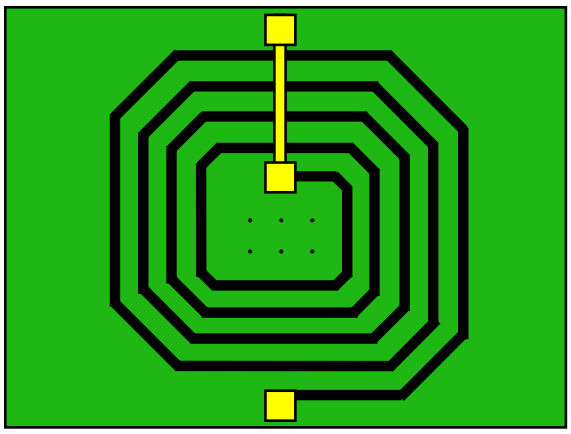


Figure 8.18 Spiral configuration of an inductor in IC chip

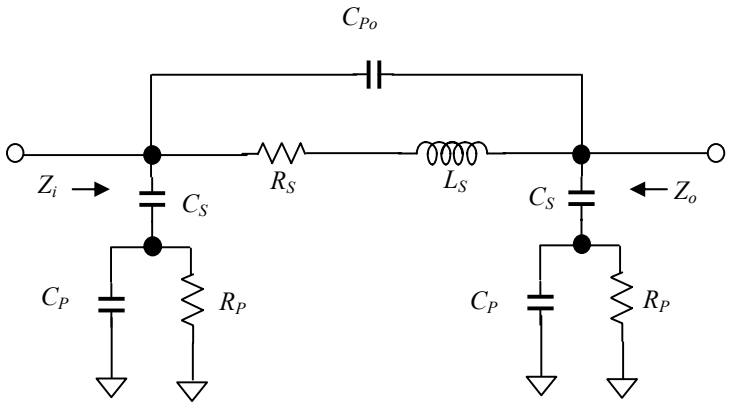


Figure 8.19 Model of a spiral inductor

Compared to other parts, there are two difficulties to a spiral inductor:

- Large area: Its dimension of size is usually over 100 μm^2. It is more than ten or hundred times than that of a resistor. The large area impacts the cost of IC chip directly since the cost of IC chip is directly proportional to its die area.

- Extremely low Q value: Up to most announces, the Q value of a spiral inductor in an RFIC is less than 10. In an RF design with discrete parts, the Q value of inductors generally is over 100. For example, the Q value of a chip inductor is around 120.

The inductor is one of the indispensable parts of an RF circuit design because it is a phase-shift part. In past years, some scientists and engineers have attempted to avoid in the use of inductors in certain RF circuit designs such as the filter design. For instance, the log-domain filtering scheme is one of the technologies in constructing of a filter by only device and capacitor. Unfortunately, it is only realistic when the operating frequency is low, say, below 1 GHz. Secondly, it is an active circuit so that the DC current or power consumption must be paid. Similar to the log-domain filtering technology, much effort has been put on the development of the equivalent inductor circuit, the gyrator. The restriction or jeopardy is the same as in the log-domain filtering technology.

How come the Q value of a spiral inductor is so low? Much hard research in last decades has been conducted up to date. The reasons can be outlined in the following sub-sections.

6.3.1 Skin Effect

Skin Effect tells us that the electromagnetic wave tends to be propagated along a conductor's surface or the conductor's "skin". The strength of the electromagnetic wave is attenuated as the depth from the surface of the conductor is increased. The depth of the Skin Effect, δ, is defined as a depth where the strength of the electromagnetic wave is attenuated to 1/e or 37% of its strength at the surface, and is a function of the frequency f, the conductivity σ, the magnetic permeability μ, and the electric permittivity ε.

$$\delta = \frac{1}{\omega\sqrt{\frac{\mu\varepsilon}{2}\left[\sqrt{1+\frac{\sigma^2}{\omega^2\varepsilon^2}}-1\right]}} \approx \sqrt{\frac{2}{\omega\mu\sigma}} \quad , \tag{8.5}$$

$$\varepsilon = \varepsilon_r \varepsilon_v \quad , \tag{8.6}$$

$$\mu = \mu_r \mu_v \quad , \tag{8.7}$$

where $\omega = 2\pi f$ = Angular frequency;
f = frequency;
ε_r = relative electric permittivity of the metal,
ε_v = Electric permittivity in vacuum or in free space,
μ_r = Relative magnetic permeability of the metal,
μ_v = Magnetic permeability in vacuum or in free space, and

$$\varepsilon_v = 8.854 * 10^{-12} \approx \frac{1}{36\pi * 10^9} \text{, f/m,} \tag{8.8}$$

$$\mu_v \approx 4\pi * 10^{-7} \text{ , H/m,} \tag{8.9}$$

For copper,

$$\sigma = 5.8 \text{ x } 10^7, \Omega^{-1}/\text{m,} \tag{8.10}$$

If

$$\sigma >> \omega \varepsilon \text{ ,} \tag{8.11}$$

then we can calculate the depth of skin effect, δ, as shown in the Table 8.7.

Table 8.7 Variation of depth of skin effect vs. frequency for copper.

Operating frequency	100M	1G	5G	10G	Hz
Depth of Skin Effect	6.61	2.09	0.93	0.66	μm

It is well known that in the transportation of an electromagnetic wave, it is somewhat a waste of material if the thickness of the metal is greater than the depth of skin effect. On the contrary, it causes power loss of the electromagnetic wave if the thickness of the metal is less than the depth of skin effect. The power loss is equivalent to the low Q value for an inductor.

In the case of integrated circuit, the thickness of metal layer is in the order of

$$T \sim 1 \text{ μm.} \tag{8.12}$$

Referring to Table 8.7, one might conclude that the low Q value of the inductor on an IC chip is due to the fact that the metal layer is too thin when the operating frequency is below 1 GHz, and that the low Q value of inductors on an IC chip could be extended when the operating frequency is over 1 GHz and up to 10 GHz, though the thickness of the metal layer is comparable with the depth of skin effect. On the basis of such a conclusion, many experiments have been conducted to enhance the Q value of IC inductors by means of increasing of thickness of the spiral inductor's wire. Unfortunately, the Q value of IC inductors is not significantly increased even though thickness of the metal wire is piled up to several micro meters. It implies that the thin thickness of the metal layer is not the main reason to bring about the low Q value of the IC inductor.

8.3.2 Attenuation due to Substrate

The metal winding of an IC spiral inductor is laid on a substrate. The electric permittivity of the substrate, ε_s, is generally much higher than the permittivity of the free space, so that the electromagnetic field as well as its power in the substrate is much stronger than in the free space. The electromagnetic field can be stored, propagated, and consequently

attenuated in the substrate. Then, the substrate might be the main objective to cause the low value of the IC spiral inductor.

An outstanding experiment has been conducted in UCLA, US. The substrate beneath the spiral inductor is dug away, which can be explained by a draft as shown in Figure 8.20. 4 bars are constructed to support the spiral wires. This is a difficult and high technological task. Professors and students, scientists and engineers in UCLA have been working hard and finally got this special treatment done successfully. The Q value of this spiral inductor is increased but unfortunately, it is still not high enough to satisfy most RFIC designs. In the UCLA, some other projects about the enhancement of the Q value for the IC spiral inductors have been conducted for years.

Again, it implies that the main reason to the low Q value of the IC spiral inductor is not the attenuation due to the substrate.

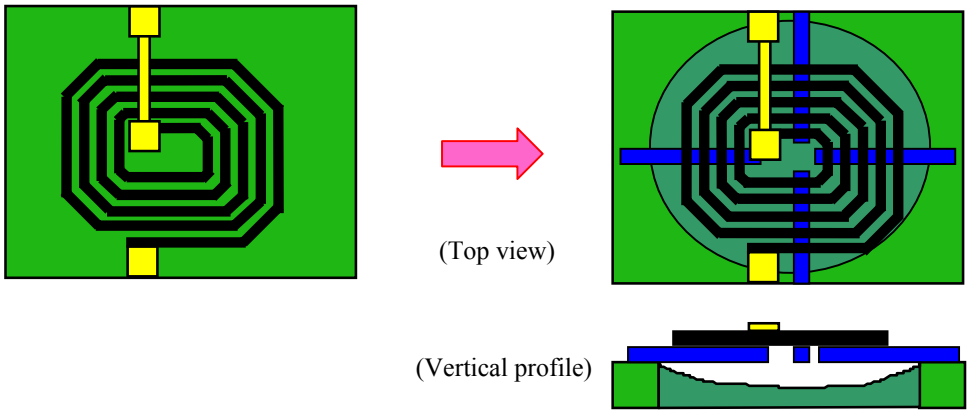
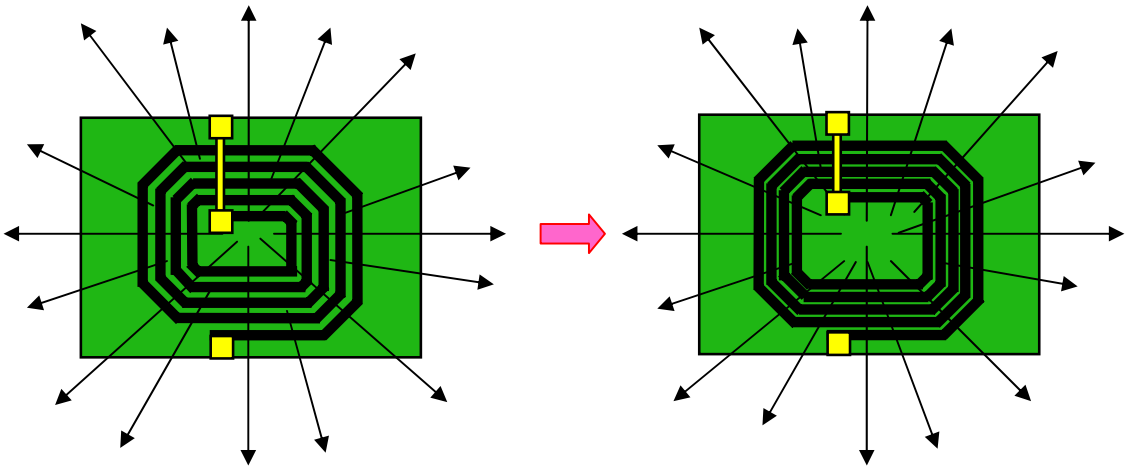


Figure 8.20 Digging of the substrate beneath the spiral wire away

8.3.3 Flux Leakage

There is no doubt that the leakage of flux is one of the factors which causes the low Q value of IC spiral inductors. In Figure 8.21 it is shown that the flux leakage is from the spacing or gap between the windings. The leakage is reduced when the spacing or gap is squeezed. The flux leakage shown in Figure 8.21 (b) is less than that as shown in Figure 8.21(a).

The question is, what is the degree of this factor to impact the Q value down from the case without flux leakage? The experiment illustrates that the Q value is not improved too much even if the spacing between the windings is squeezed to an allowed tolerance value. For example, the Q value might be increased from 5 to 6 when the spacing between the windings is squeezed from 10 μm to 1 μm.



(a) (b)

Figure 8.21 Squeezing of windings so as to reduce the spacing between windings.

Well, another type of flux leakage might be the magnetic flux emitted from the winding to the sky. Theoretically, power exists in any space where flux exists. It is therefore imagined that the Q value of the IC inductor might be improved when the flux is confined within a "Sandwich" as shown in Figure 8.22. Above the windings of an inductor a cavity is formed by a kind of magnetic material with high permeability μ. As long as the permeability of the Sandwich material is high enough, very less flux can escape from the Sandwich cavity. Consequently, the Q value is expected to be high. Yes, experiment did demonstrate the improvement of the Q value. Unfortunately, like those results in other experiments mentioned above, the improvement is quite poor.

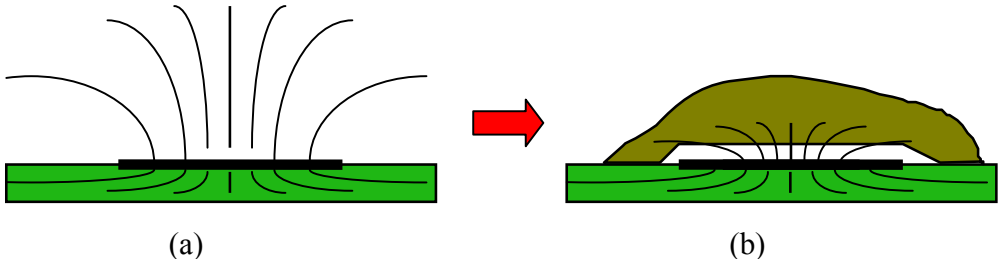


Figure 8.22 Sandwich the IC spiral inductor so as to confine the flux within the "Sandwich"

It is therefore concluded again that the low Q value is not mainly due to the flux leakage.

8.3.4 Flux Cancellation

In order to find out the main reason of the low Q value for an IC spiral inductor, engineers and scientists worked very hard for many years. All the experiments discussed above failed to find the main reason of the low Q value for an IC spiral inductor. It is, in fact, due to the inherent drawback of the spiral configuration: the cancellation of the flux between the windings. As shown in Figure 8.23, in the spacing between two windings, the flux produced by the inner winding is cancelled by the flux produced by the outer winding. The flux cancellation is more important than the flux leakage to the reduction of Q value.

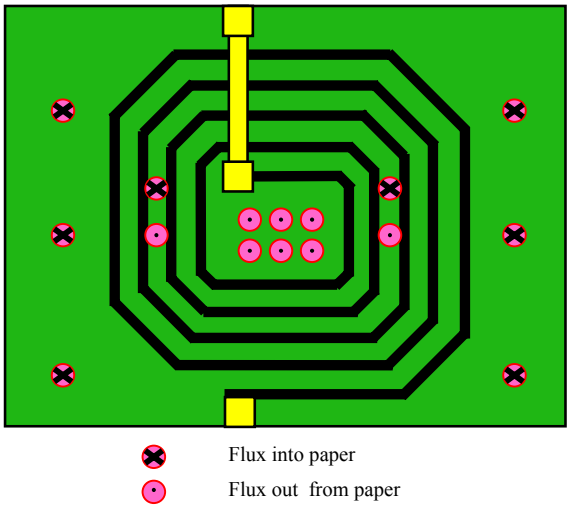


Figure 8.23 Cancellation of flux between two windings

Let's recall the well-known Farady's law. It says that a electromotive force or an induced voltage will be produced when an alternative driving current is applied to an ideal inductor. It can be expressed as

$$\varepsilon = -L\frac{di}{dt}, \tag{8.13}$$

where ε = Electromotive force or induced voltage,
i = Alternative driving current, and
L = Inductance of the inductor.

An ideal inductor means that there is neither flux leakage nor flux cancellation. As shown in Figure 8.24, all of flux are going through all the windings simultaneously.

In this ideal case, L is a real number.

In an actual IC spiral inductor with flux leakage and flux cancellation, the Farady Law (8.13) must be modified such as

$$\varepsilon = -L\frac{di}{dt} - ri\ , \tag{8.14}$$

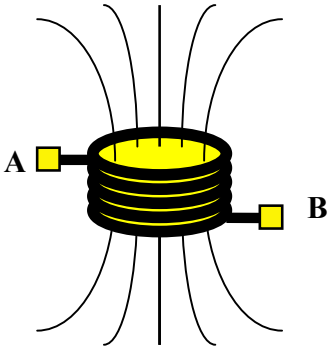


Figure 8.24 Farady's law with an ideal inductor

where the additional term, *ri*, represents the voltage drop due to the ohmic resistance, which is equivalent to the flux leakage and flux cancellation.

The ohmic resistance can be expressed by the Q value of the inductor. Up to the definition of Q value,

$$r = \frac{L\ \omega}{Q}\ , \tag{8.15}$$

where ω is the angular frequency of the alternative driving current, which can be expressed by

$$i = i_o e^{j\omega t}\ . \tag{8.16}$$

From (8.14), (8.15), and (8.16), we have

$$\varepsilon = -L\frac{di}{dt} - \frac{L\omega}{Q}i = -(L - j\frac{1}{Q})\frac{di}{dt} = -L_{eff}\frac{di}{dt}, \tag{8.17}$$

where

$$L_{eff} = L - j\frac{1}{Q}\cdot \tag{8.18}$$

It is found that L_{eff} is a complex number. In an ideal inductor without flux leakage and flux cancellation, Q is infinitive and then L_{eff} approaches its ideal value L. In other words, a complex inductance value L_{eff} represents the Q value of the inductor is not infinitive but a limited value.

The flux leakage and flux cancellation in a practical IC spiral inductor is the main reason of its low Q. This jeopardy comes from its inherent spiral configuration. One of possible solutions, of course, is to change its spiral configuration. There are many metallic layers in advanced IC processes. The flux leakage and flux cancellation might be eliminated when the IC inductor is built by multiple of metallic layers. Figure 8.25 shows the primary idea. There are 4 metallic layers. Every layer contains one winding of the inductor. All of windings are vertically overlapped together. A conductive via (hole) connects two adjacent windings together. By such a configuration, the flux leakage and flux cancellation might be significantly reduced down. Therefore a new inductor with a high Q value is expected.

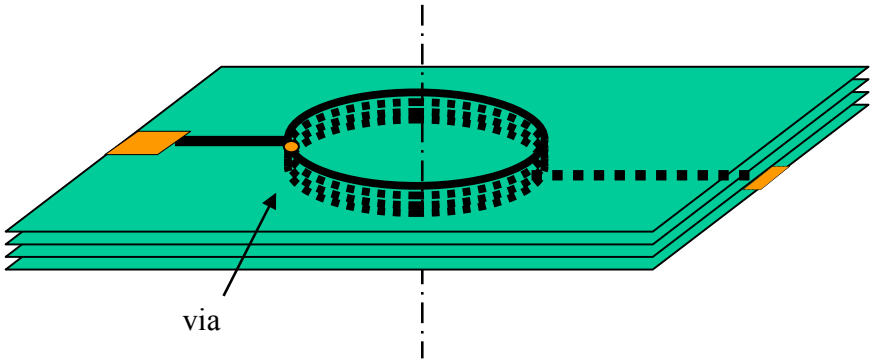


Figure 8.25 Build an IC inductor by multiple of metallic layers

Unfortunately, the test result indicates the Q value of such an inductor is even worse than the spiral one. This is due to the significant voltage drop and attenuation when the alternative current goes through the vias (holes) between metallic layers. Figure 8.26 shows an equivalent circuit of a small via (hole) from one metallic layer to the other. It consists 4 parts. *L* in series with *r* represents a lossy inductor. The *r* increases as the diameter of the via is reduced. The diameter of the via is about couple micro meters but the *r* is already tens or hundreds of ohms. It is therefore a very lossy inductor. In addition to this lossy inductor, there are two capacitors, C_1 and C_2, represents the wall and fringe capacitance of the via. The high *r* and these two capacitors are a disaster bottleneck for this type of inductor configuration.

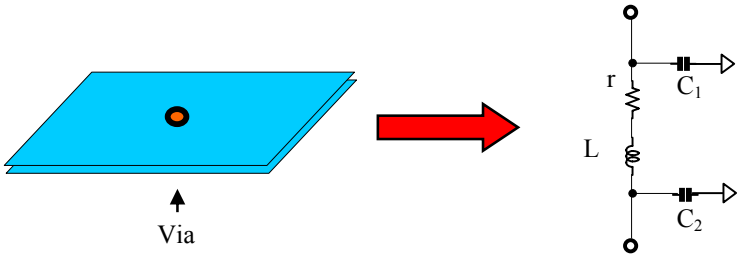


Figure 8.26 Equivalent circuit of a small via (hole).

8.3.5 A Possible Solution --- Negative Resistance Compensation

It seems very difficult to enhance the Q value of the IC spiral inductor in terms of the configuration improvement. Negative resistance compensation might be one of issues to solve the jeopardy. As a matter of fact, a low Q inductor can be represented by an ideal inductor in series with a resistor *r*, which is the representation of low Q value of the inductor. Figure 8.27 shows its equivalent.

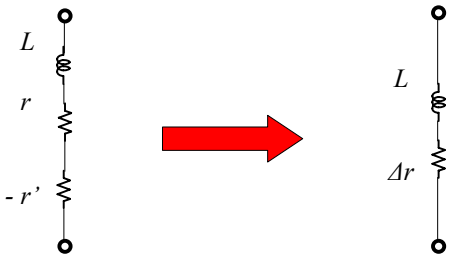


Figure 8.27 Equivalent circuit of an inductor with low Q by a resistor

Now assuming that a negative resistor, -r', can be created and is in series with the low Q inductor. The resulted resistance and Q value can be expressed by

$$\Delta r = r - r' \quad , \tag{8.19}$$

and,

$$Q = \frac{L\omega}{r - r'} = \frac{L\omega}{\Delta r} \quad . \tag{8.20}$$

Theoretically, the Q value could be compensated to infinity if the negative resistance, -r', is equal to the original resistance of inductor, r, in magnitude. However, in engineering design, it brings many new problems. There are 3 barriers which must be overcome:

1) Bandwidth control:

Over the operating bandwidth the resistance of the inductor must be compensated so that the resultant resistance, Δr, must be positive as shown in Figure 8.28. Otherwise, it could introduce the circuit to an oscillation state within the desired bandwidth. On other hand, outside the operating bandwidth, the negative resistance "generator" must produce no negative resistance so as to guarantee the oscillation does not happen outside the bandwidth.

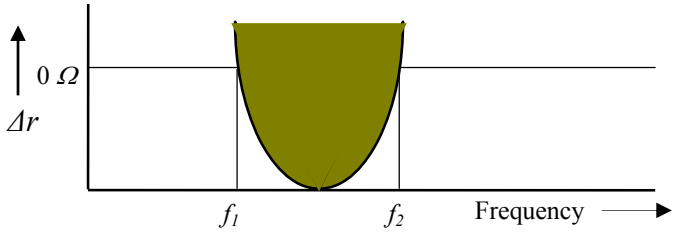


Figure 8.28 The negative resistance must be controlled.

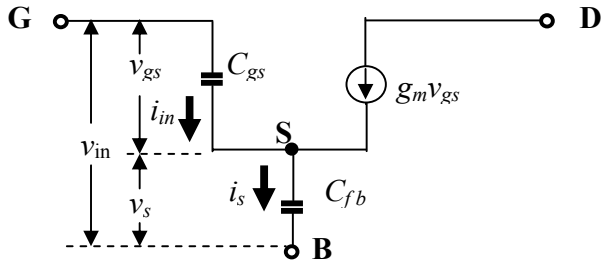
2) Current drain:
The negative resistance is usually produced in terms of an active circuit. The current drain requirement is one of the important factors to be considered in the enhancement of Q value. For example, if 1 mA of current drain needs to be provided for the negative resistance compensation for one inductor, then 10 mA of current must be spent on that for 10 inductors. This might be higher than the stand-by current for all the system. Reduction of the current drain required is an important engineering objective.

3) Noise:
There is no noise for an ideal inductor. However, an active circuit for the negative resistance inevitably brings about a lot of noise. There is no way to get rid of the additional noise, but just try to reduce it as much as possible.

Let's introduce two basic means in the creation of negative resisitance.

8.3.5.1 Negative Resistance Generator with a FET

A FET operating in low current drain with a negative bias can serve as a negative resistance generator. Its equivalent model could be simplified as shown in Figure 8.29.



C_{gs} : Gate-source capacitance
C_{fb} : Source-substrate capacitance
g_m : Transconductance
v_{gs} : Gate-source voltage
v_s : Source-substrate voltage
v_{in} : Gate-substrate voltage
i_{in} : Input current
i_s : Source-substrate current

Figure 8.29 The equivalent circuit when a FET serves as a negative resistance generator.

It is easily to find that

$$v_{gs} = \frac{i_{in}}{j\omega C_{gs}}, \tag{8.21}$$

$$v_s = \frac{i_s}{j\omega C_{fb}}, \tag{8.22}$$

$$i_s = i_{in} + g_m v_{gs} = i_{in}\left(1 + \frac{g_m}{j\omega C_{gs}}\right), \tag{8.23}$$

$$v_{in} = v_{gs} + v_s = i_{in}\left(\frac{1}{j\omega C_{gs}} + \frac{1}{j\omega C_{fb}} + \frac{g_m}{j\omega C_{gs} j\omega C_{fb}}\right), \quad (8.24)$$

$$z_{in} = \frac{v_{in}}{i_{in}} = \frac{1}{j\omega\left[\frac{C_{gs}C_{fb}}{C_{gs}+C_{fb}}\right]} - \frac{g_m}{\omega^2 C_{gs}C_{fb}}. \quad (8.25)$$

Or,

$$z_{in} = r_{in} + jx_{in}, \quad (8.26)$$

$$x_{in} = -\frac{1}{\omega\left[\frac{C_{gs}C_{fb}}{C_{gs}+C_{fb}}\right]}, \quad (8.27)$$

$$r_{in} = -\frac{g_m}{\omega^2 C_{gs} C_{fb}}. \quad (8.28)$$

The drawback of this generator is that the negative resisitance, as shown in expression (8.28), is very sensitive to the frequency. Consequently it is very hard to control the negative resistance either within or outside the operating badwidth.

8.3.5.2 Negative resistance generator with transformer

Figure 8.30 shows an improved negative resistance generator, which mainly consists of not only a FET but also a transformer. One half of the transformer, L_1, is functioning as the IC inductor, and another half of the transformer, L_2, is functioning as the negative resistance generator for L_1. The capacitor C is for the 90° phase shift between i_1 and V_{gs}. The resistors R_1 and R_2 are for the adjustment of the feedback.

Assuming that

$$V_{gs} = V_{gso} e^{-j\omega t}, \quad (8.29)$$

$$i_1 = C\frac{dV_{gs}}{dt} = -j\omega C V_{gs}, \quad (8.30)$$

$$i_2 = g_m V_{gs}, \quad (8.31)$$

$$\frac{i_2}{i_1} = \frac{g_m}{-j\omega C}, \quad (8.32)$$

$$Z_{in} = R_1 + j\omega M \frac{i_2}{i_1} + j\omega L_1 = (R_1 - \frac{g_m M}{C}) + j\omega L_1 \quad . \qquad (8.33)$$

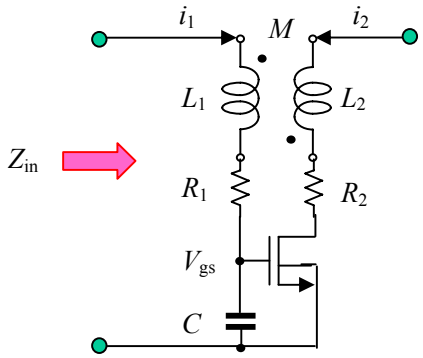


Figure 8.30 The equivalent circuit when a FET serves as a negative resistance generator with a transformer.

8.4 Layout

The layout is an important part of the whole RF or RFIC design process. Most layout problems in a practical RF or RFIC design come from incorrect runners and the incorrect allocation of parts.

8.4.1 Runners

Runners connect two parts together.

In either RF or RFIC circuits, a signal runner is a micro strip line. A long runner is therefore a long micro strip line and might be more important than a capacitor, an inductor, or a resistor in the performance of the circuit. In some RF designs, the circuit mainly consists of only the device and micro strip lines. Strictly handling of the runners is always the guarantee of a successful RF and RFIC design.

Let's examine the characterization of the micro strip line in the RFIC design. Figure 8.31 denotes the various parameters for a micro strip line. In a CMOS process, with silicon substrate, a micro strip line can be described by the distribution parameters, the capacitance per unit length in respect to the substrate, C_{msl}, and the self-inductance per unit length along the runner, L_{msl}.

$$C_{msl} = \varepsilon_{ox}\left[2.42 + \frac{w}{X_{\text{int}}} - 0.44\frac{X_{\text{int}}}{w} + \left(1 - \frac{X_{\text{int}}}{w}\right)^6\right], \text{ F/cm,} \qquad (8.34)$$

where ε_{ox} is the electric permittivity of the silicon-oxide layer, and

$$\varepsilon_{ox} = 3.45\text{x}10^{-13}, \text{ F/cm}, \qquad (8.35)$$

$$L_{msl} = 2\ln\left[\frac{8X}{w} + \frac{w}{4X}\right], \text{ nH/cm}, \qquad (8.36)$$

where

$$X = X_{\text{int}} + X_{si}, \qquad (8.37)$$

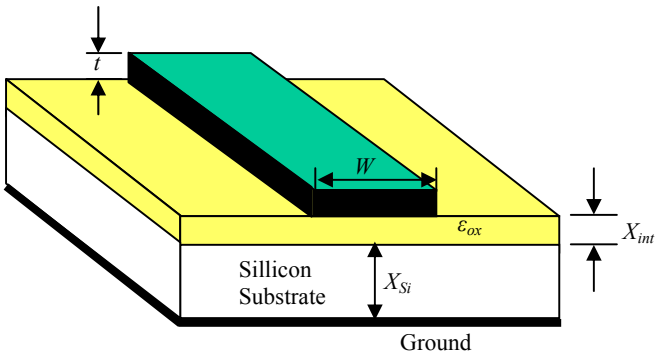


Figure 8.31 Various parameters of a micro strip line on silicon substrate.

The quarter wavelength is

$$\frac{\lambda}{4} = \frac{1}{4f} = \frac{2\pi}{4}\sqrt{C_{msl}\,L_{msl}}\,, \qquad (8.38)$$

Table 8.8 An example of quarter wavelength (*QWL*) on an RFIC die in a CMOS process

W, µm	C, pF/cm	L, nH/cm	Freq, Hz	QWL, mm	l,(5°), µm	Freq, Hz	QWL, mm	l,(5°), µm	Freq, Hz	QWL, mm	l,(5°), µm	Freq, Hz	QWL, mm	l,(5°), µm
1	1.45	16.59	1.0E+09	16..09	894	2.4E+09	6.71	373	5.8E+09	2.77	154	10.0E+09	1.61	89
2	2.24	15.20	1.0E+09	13.55	753	2.4E+09	5.65	314	5.8E+09	2.34	130	10.0E+09	1.36	75
3	3.00	14.39	1.0E+09	12.04	669	2.4E+09	5.02	279	5.8E+09	2.08	115	10.0E+09	1.20	67
4	3.73	13.82	1.0E+09	11.01	612	2.4E+09	4.59	255	5.8E+09	1.90	105	10.0E+09	1.10	61
5	4.45	13.37	1.0E+09	10.25	569	2.4E+09	4.27	237	5.8E+09	1.77	98	10.0E+09	1.02	57
10	7.98	11.98	1.0E+09	8.08	449	2.4E+09	3.37	187	5.8E+09	1.39	77	10.0E+09	0.81	45
15	11.46	11.17	1.0E+09	6.99	388	2.4E+09	2.91	162	5.8E+09	1.20	67	10.0E+09	0.70	39
20	14.93	10.60	1.0E+09	6.29	349	2.4E+09	2.62	145	5.8E+09	1.08	60	10.0E+09	0.63	35
30	21.84	9.79	1.0E+09	5.41	300	2.4E+09	2.25	125	5.8E+09	0.93	52	10.0E+09	0.54	30
40	28.75	9.21	1.0E+09	4.86	270	2.4E+09	2.02	112	5.8E+09	0.84	47	10.0E+09	0.49	27
50	35.66	8.77	1.0E+09	4.47	248	2.4E+09	1.86	104	5.8E+09	0.77	43	10.0E+09	0.45	25
60	42.56	8.40	1.0E+09	4.18	232	2.4E+09	1.74	97	5.8E+09	0.72	40	10.0E+09	0.42	23
75	52.92	7.96	1.0E+09	3.85	214	2.4E+09	1.61	89	5.8E+09	0.66	37	10.0E+09	0.39	21
100	70.17	7.38	1.0E+09	3.47	193	2.4E+09	1.45	80	5.8E+09	0.60	33	10.0E+09	0.35	19
200	139.17	6.00	1.0E+09	2.74	152	2.4E+09	1.14	63	5.8E+09	0.47	26	10.0E+09	0.27	15
500	346.18	4.22	1.0E+09	2.07	115	2.4E+09	0.86	48	5.8E+09	0.36	20	10.0E+09	0.21	11

Note: l (5°), µm = Length of runner corresponding to phase shift 5°.

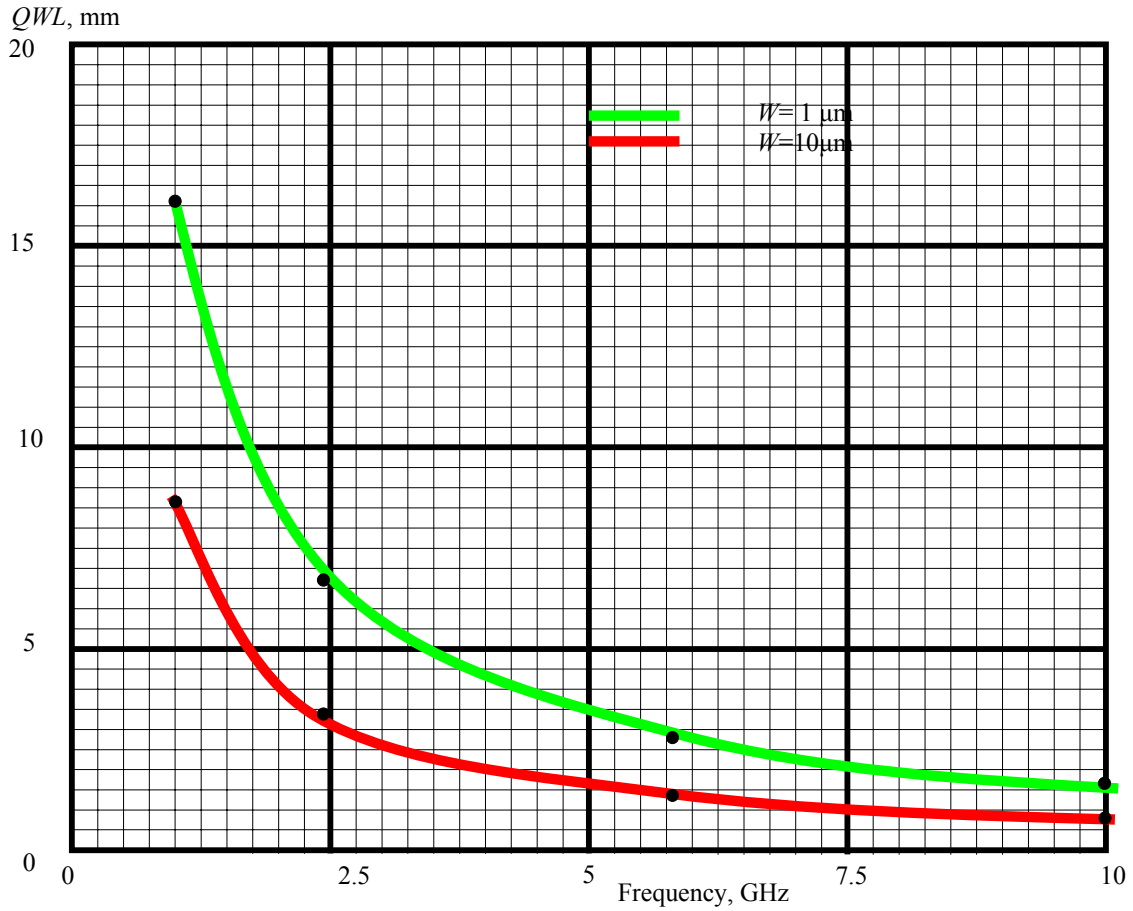


Figure 8.32 Quarter wavelength (QWL) vs. Frequency

As an example, assuming that X_{int}=0.5 μm, X_{si}=500 μm, and ε_{ox}= 3.45x10^{-13} F/cm, the quarter wavelength with different frequency and the width of the micro strip line are listed in Table 8.8. Figure 8.32 plots two curves of quarter wavelength vs. frequency when the width of the micro strip line is 1 μm and 10 μm respectively. The quarter wavelength is a very important parameter in the layout design. An RF signal on a runner with quarter wavelength could experience short circuit to open circuit, or zero impedance to infinitive impedance. In the layout work, the runner is better kept as short as possible, that is,

$$l << \frac{1}{4}\lambda . \qquad (8.39)$$

If the length of the runner is comparable or greater than the quarter wavelength, the runner must be treated as a part in the circuit simulation. In the practical RFIC design, the width of the runner typically is taken in the range of

$$100\ \mu m > w > 5\ \mu m. \qquad (8.40)$$

As a practice, let's examine the effect of a runner on the impedance matching. Assuming that the impedance at point P in the Smith Chart of Figure 8.33 is the original load impedance and is desired to be matched to the source impedance at point O.

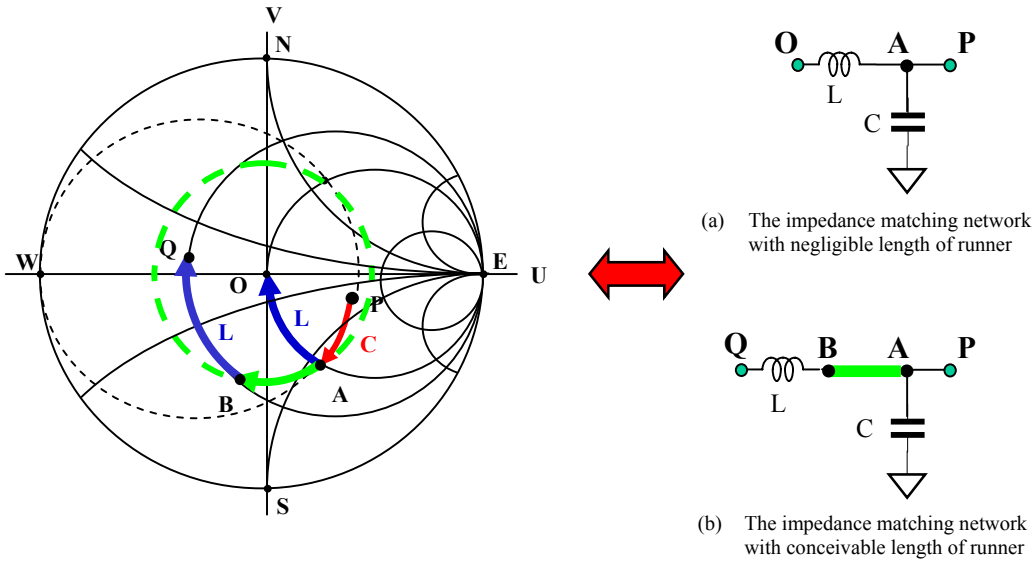


Figure 8.33 Impedance matching is degraded by the runner from A to B.

The first case is shown in Figure 8.33 (a), in which the length of runner in the impedance matching network is negligible. On the Smith Chart, the impedance of the original load shifts from P to A after a capacitor *C* is connected to the original load in parallel. The value of the capacitor can be measured and calculated from Smith Chart so that the point A is exactly located on the resistance circle which passes the source impedance at point O.

Then, the impedance is moved from A to O if an inductor L is connected to the point A. The value of the inductor can be measured and calculated from Smith Chart.
The second case is shown in Figure 8.33 (b), in which the length of runner in the impedance matching network is conceivable, which corresponding the impedance moved from A to B on Smith Chart. The arc AB depends on the length and width of the runner. Then, the impedance is moved from B to Q if the same inductor, L, is connected to the point B.

Owing to the runner from A to B, the final impedance drifted from O to Q. The drifted amount depends on the length and width of the runner AB. For example, if the runner AB is $\lambda/16$, or 45^{o}, and if the width of runner is 10μm, then from Table 8.7, we can find the length of the runner, $\lambda/16$, is

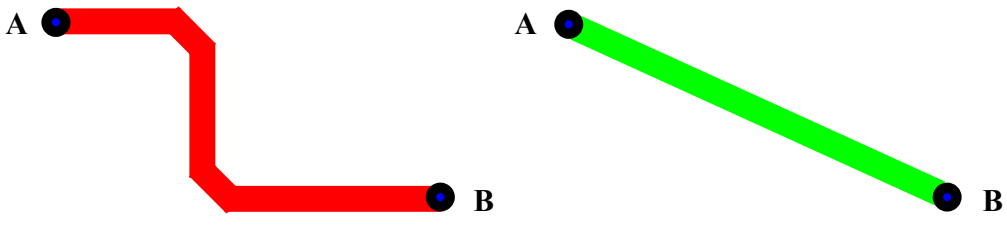
2.02 *mm* for 1*GHz* & w=10*μm*,
or, 0.84 *mm* for 2.4*GHz* & w=10*μm*,
or, 0.35 *mm* for 5.8*GHz* & w=10*μm*,
or, 0.20 *mm* for 10 *GHz* & w=10*μm*,

which is dependent of the operating frequency.

- **Style of runner**

Many new engineers pursue the runner in a "nice-looking" style. Very often they draw the runner along with either East-West or South-North direction as shown in Figure 8.32 (a). Like the layout in the digital circuit design, a bunch of runners are in parallel together. It is really nice-looking, just like many well-trained soldiers parading on a street.

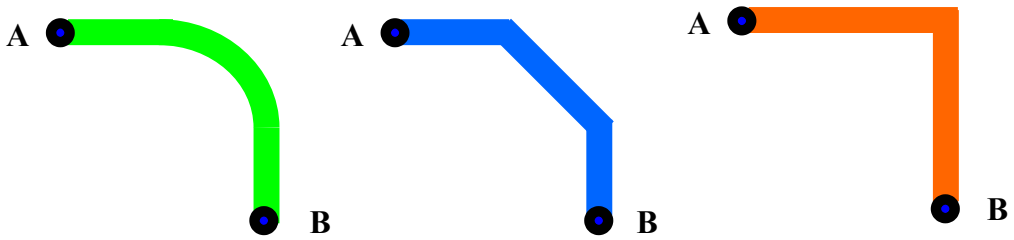
However, such a kind of runner style is very bad to the performance. Not only it increases the length of runner, but also it introduces very strong cross-talk among that bunch of runners. In addition, the runners in parallel induce mutual capacitance and inductance. Figure 8.34 shows two runner styles drawing from A to B. The style (a) is a bad style whereas the style (b) is good one. The runners must be kept as short as possible.



(a) "Nice looking" and "E-W, S-N" orientation (b) As short as possible -Yes!

Figure 8.34 Two runner styles from A to B.

- **Corner of the runner**



(a) Circular-the best! (b) 45°-OK! (c) Rectangular- Lousy!

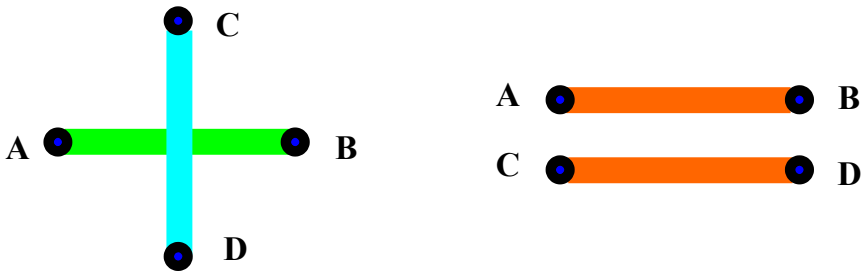
Figure 8.35 Three different corners of runner from A to B.

Figure 8.35 shows three different corners of runners. The corner (a) is much better than (b) and (c) because

- It is smooth.
  Any angle, especially any sharp angle of corner makes the singular point to the electromagnetic field and introduces fair amount of radiation.
- It is shortest connection.

- **Preference to draw adjacent runners**

First, we prefer to draw two adjacent runners in perpendicular, and try to not in parallel as possible. Figure 8.36 illustrates such a preference.



(a) In perpendicular - Welcome! (b) In parallel - Reluctant!

Figure 8.36 Draw two adjacent runners in perpendicular, not in parallel

Very often we cannot but draw two adjacent runners in parallel as shown in Figure 8.37. In such a case, the spacing between the two runners must be 3 times greater than the width of the runners, that is,

$$S \geq 3W \quad , \tag{8.41}$$

so that the cross-talk can be reduced to an acceptable level. However, it does not matter if the two adjacent runners are transporting a DC voltage or current.

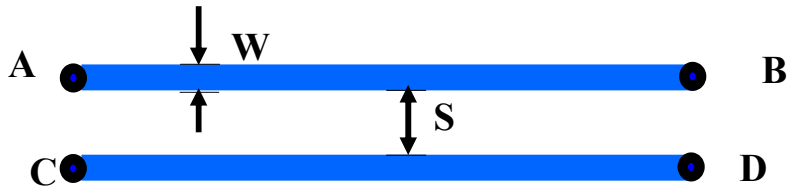
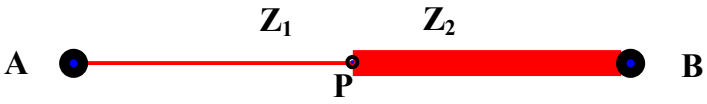


Figure 8.37 Spacing between two runners in parallel.

- **Smooth of a runner**

Smoothness is important not only for the corner of a runner, but also important for the entire runner. Figure 8.38 (a) shows the width of a runner suffers a suddenly change. It is well known that the characteristic impedance of a micro strip line is mainly determined by its width. The characteristic impedance at point P jumps from Z_1 to Z_2. Consequently it, in fact, is an impedance transformer. This additional jump of impedance might result in a disaster in the circuit performance: The RF power might be reflected back and forth at point P. In addition, at point P, the RF signal suffers radiation. On the contrary, the width of the runner in Figure 8.38 (b) is gradually changed so that the variation of impedance is smooth. The additional reflection is very weak and the radiation is very little at any point on the runner.



(a) Width of a runner changed suddenly

(b) Width of a runner changed gradually

Figure 8.38 Width of runner changed gradually, not suddenly.

- **Summary**

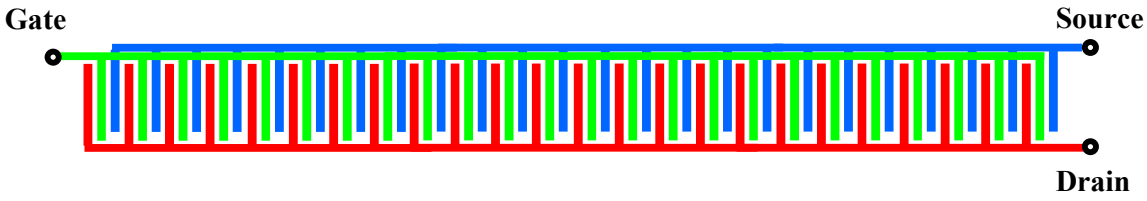
The discussion about runner above can be outlined to three points:

1) As short as possible,
2) As smooth as possible,
3) As perpendicular from each other as possible.

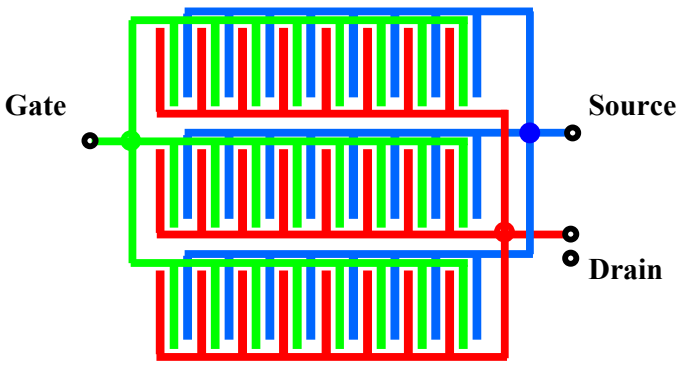
8.4.2 Parts

- **Device**

In a practical RF design, a device may consist of many sub-devices. For example, a MOSFET device is a combination of many "fingers" of the source, gate, and drain, as shown in Figure 8.39. There are two combination styles of sub-devices. One is all of devices being lined up as a long "dragon" as shown in Figure 8.39 (a). It is not a good combination way. There are three connected points, source, gate, and drain, to be connected to other parts in the circuit. The current from the fingers flowing to the connected point are not simultaneous. The current flowing to the connected point from the nearest finger spends the least time



(a) Device is lined up as a "dragon".



(b) Device is arranged in a "square" shape.

Figure 8.39 Devices must be arranged in a "square" but not a "dragon" shape.

whereas the current flowing to the connected point from the farthest finger spends longest time. A phase difference of an RF current within the device produces additional noise. In the extreme case if the time difference is comparable with the period of RF signal, the device cannot be operated at that operating frequency.

Figure 8.39 (b) shows a better style to adapt in the RF layout. The difference of current arrival time from the sub-device to the connected point is approached to a minimum since all of sub-devices are arranged in a "square" shape.

For the sub-devices arrangement, it is therefore concluded: Not "dragon", but "square"!

- **Inductor**

In the IC layout for an inductor, in order to prevent from cross-talk between adjacent inductors via substrate, the designer must pay attention on two issues:

1) The guard ring must be added and circled to every inductor. Usually a guard ring consists of two wells: N-well and P-well. N-well is connected to the V_{dd} and P-well is connected to the ground so that it forms a reverse-bias between the P-N junctions.
2) There must be ample space between two inductors.

Figure 8.40 illustrates there is possible cross-talk between the inductors through the flux interference in the air, though a guard ring is available for each inductor. In order to reduce the air cross-talk, in a practical RF layout, it requires that

$$D > d_1, \tag{8.42}$$

$$D > d_2. \tag{8.43}$$

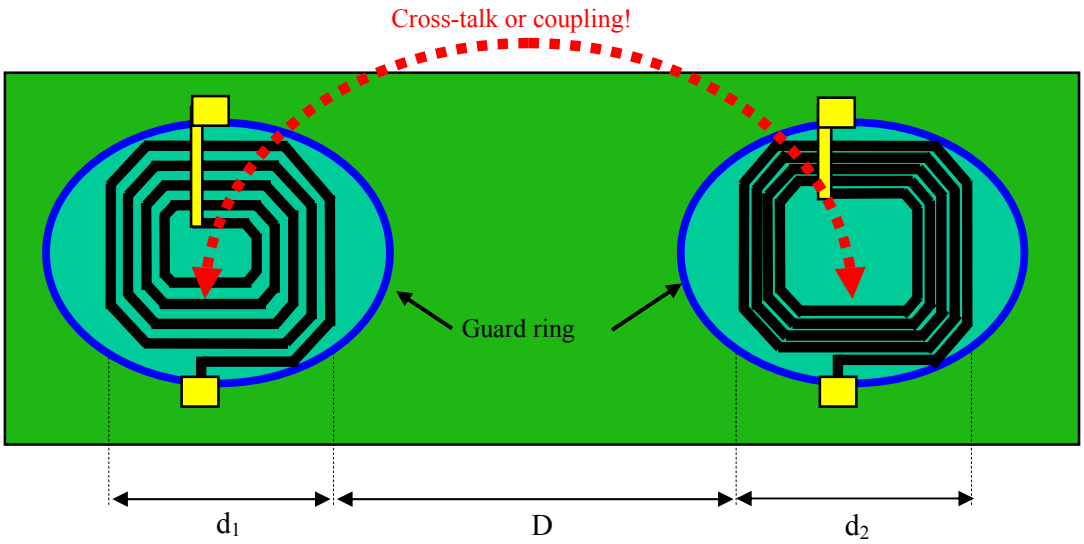


Figure 8.40 Cross-talk between two inductors

- **Resistor**

The relative tolerance of an IC resistor is highest among the parts. It could reach to 20%. In order to lower the tolerance, a combination of resistors in parallel is applied. For example, Figure 8.41 shows that a desired 1 kΩ resistor is replaced by a combination of 10 resistors in parallel, the value of each resistor being 10 kΩ.

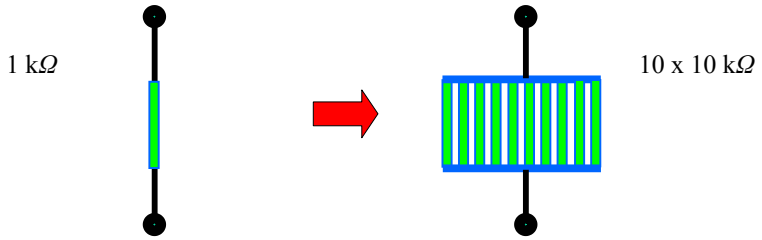


Figure 8.41 A desired 1 kΩ resistor is replaced by 10 resistors

- **Capacitor**

Besides IC spiral inductors, an IC capacitor also occupies a large area on the IC die. It is therefore that in most design centers a rule is put on the RFIC circuit design: The maximum value of the IC capacitor should not exceed over 20 pF. Should a capacitor with the value greater than 20 pF is required, it will be removed from IC die and put on the external circuit. Studies of capacitors with high capacitor per area are in progress. The restriction of 20 pF could be removed if such a study or research is achieved. This would be a great step forward in the IC technology.

8.4.3 Variable Parts in RFIC

The cycle time of an RFIC design is about 6 months today. A successful RFIC normally needs 3 to 5 times of tape-outs. It implies that 1.5 to 2.5 years are needed for a new RFIC product. This is a long period indeed. Unlike a discrete RF circuit design, on an RFIC die a part cannot be replaced until next tape-out. This is one of difficulties in the RFIC design as opposed to the discrete design. So, this is desirable to reduce the tape-out number so as to shorten the developing period of a new product. A powerful means to do this is to modify most of RFIC parts to be variable. The following displays the possibility of the variable parts in RFIC design.

- Device

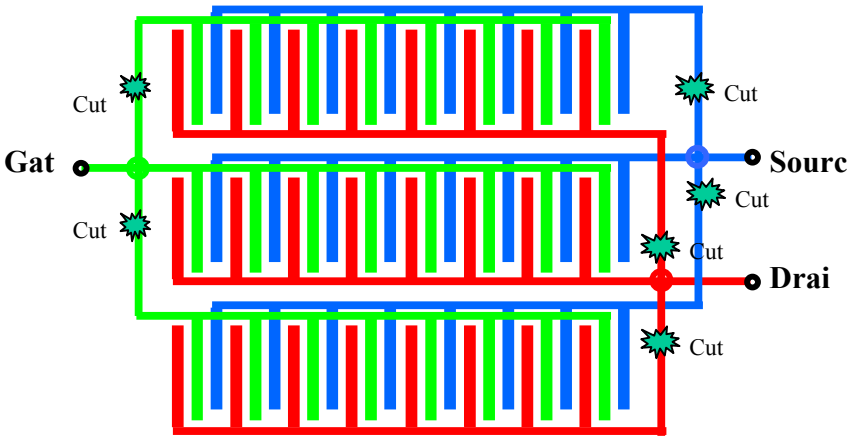


Figure 8.42 Variable device

In Figure 8.42, the device is combined with 3 sub-devices. The finger number of device can be changed if the runner between two sub-devices is trimmed and burned away. The mark "cut!" denotes that the location is to be trimmed possibly.

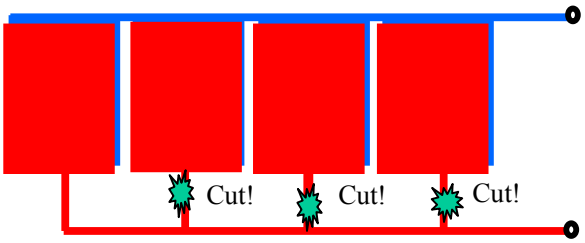
- **Other variable parts**

Figure 8.43 shows the layout for other variable parts.

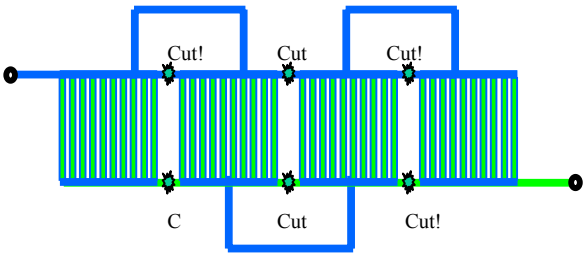
By means of the IC runner trimmer, almost all parts can be modified as a variable part. However, it is not advisable to do such a modification for the IC inductor because the spacing between its adjacent windings is very small so that in the trimming process the windings might easily be damaged.

Instead of trimming, the RF switching technology, which has been developed in recent years might be applied in the change of the IC parts from fixed to variable. All the trimmed places can be replaced by an RF switch, to which an ON/OFF control voltage is applied. It sounds quite simple. However, every ON/OFF place needs a control terminal and therefore too many control pads or pins must be added on the IC die or IC package. It makes this RF switch technology impractical or un-realistic.

(a) Variable capacitor



(b) Variable resistor



(c) Variable inductor

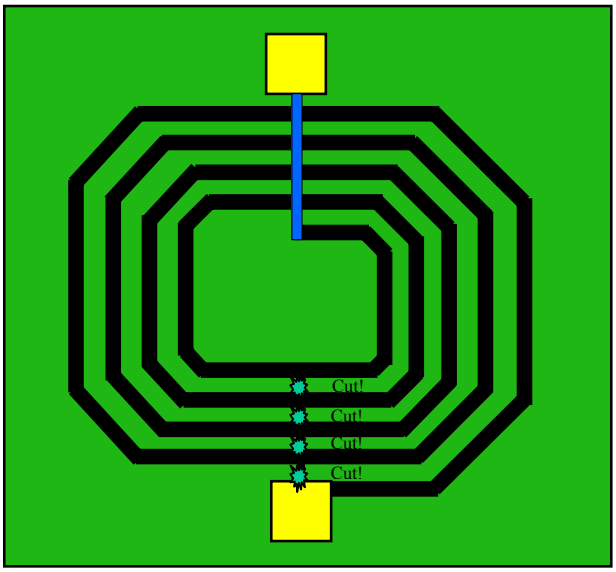


Figure 8.43 Layout of RFIC variable parts

8.4.4 Symmetry

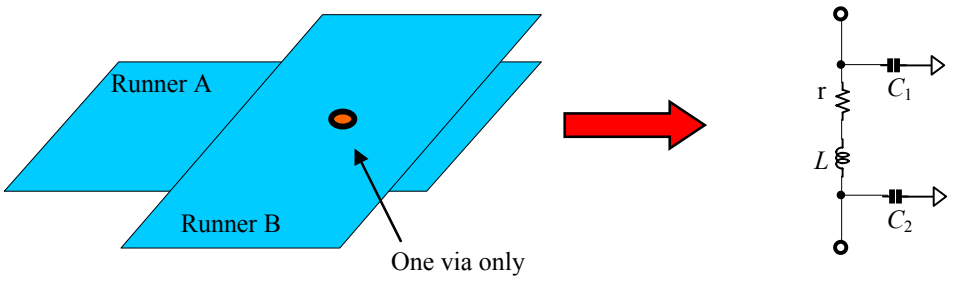
In the layout for the differential circuit, symmetry is extremely important. The asymmetry of the differential parts produces DC current offset so as to impact the circuit performance directly.

In the layout for the differential circuit, all part locations, vias, and runners, must be maintained for the geometrical symmetry.

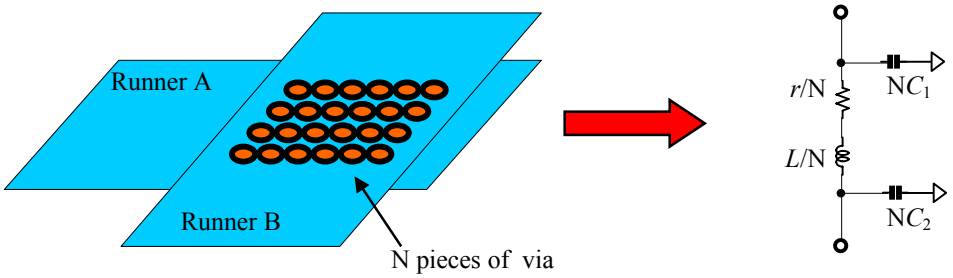
8.4.5 Via

As shown in Figure 8.44 (a), a small via produces 4 parasitic parameters: the parasitic resistance r, the parasitic inductance L, and the parasitic capacitances C_1, and C_2.

The values of r and L increase when the diameter of the via is reduced. For mechanical reasons the via is not allowed to be too large and usually its diameter is in the order of microns. As a result, the value of r and L could reach hundreds of ohms and couple nH respectively at high frequencies. In order to lower the value of r and L, many vias in parallel are put at the interception area of two runners A and B, which are supposed to be connected as shown in Figure 8.44 (b). The resulting parasitic resistance and inductance are approximately decreased N times if N vias are applied. The disadvantage is that the resulting capacitances are approximately increased N times, so that an additional attenuation of the RF signal might be suffered when the RF signal travels through this connection point.



(a) Equivalent circuit of one vias



(b) Equivalent circuit of multi-vias

Figure 8.44 Equivalent circuit of vias

8.4.6 Free Space on the Die

Many layout designers try to cover all the free space on the die with grounding metal. They think that the grounding piece is good to the isolation between parts or between blocks, or is good to help the shielding of one block from the other. The outcome, however, always has side effects: The cross-talk between blocks is enhanced and the interference between the parts is increased. The reasons are:

- The spray capacitance between the part and the grounding area is increased;
- The return current spreads over a larger grounding area, so that the cross-talk between blocks is enhanced.

So, some free space in RF circuit layout is necessary. Don't try to cover all the free space on the die by the grounding metal as possible. Also, don't squeeze or crowd the parts together.

8.5 Two Challenges in an RFIC or SOC Design

An RFIC or SOC (System On Chip) design is still quite a challenge task today though many positive achievements have been reached. Let's summarize the current status and the goals to be approached.

8.5.1 Isolation

Study for isolation between RF blocks can be divided into two stages: In the 1950's to the 1980's before the RFIC development, studies were focused on the shielding. As long as an electronic equipment contained RF blocks, a metallic box made by material with high conductivity usually covers each RF block for the good shielding. It is very heavily involved the mechanical works and the cost of the shielding boxes is sometimes more expensive than the device and all of the parts in the circuits. However, nobody dare to think about the implementing of the RF blocks with integrated circuit technology. On the other hand, in the implementing of the digital integrated circuit, the dramatic reduction of cost was very attractive and encouraging. In the 1990's, the scientists and engineers started to integrate RF circuit using the integrated technology. It was, of course, a very brave action. Up to 1995, the experiment for isolation had been achieved with a great step forward. The isolation between blocks approached 40-50 dB. RFIC technology was becoming realistic and mass production began.

There are three kinds of isolation. The first is the isolation between RF blocks. The second is the isolation between digital blocks. The third one is the isolation between RF and digital blocks. The first two have similar difficult points and have the same issue of the solution with the exception of the case when the PA block, which must be specially treated, is involved. The third one is somewhat more complicated. Normally the power of the RF block is much higher than that of the digital block since the current in the RF block is in the order of mA while the current in the digital block is μA. From the power viewpoint it looks like that a digital block is easier to be disturbed than the RF block. From the frequency viewpoint, an RF block is easier to be disturbed than a digital block because the signal in a digital block is basically a pulse. A narrow pulse is a wide band signal. It contains high frequencies, including RF components even its digital frequency is lower than the RF. Therefore, the third isolation is somehow difficult than others.

A question might be raised up to a sensitive reader: how many dB of isolation is enough for the practical purpose? In the wireless communication system, the ideal isolation level should be around 130 dB, which is 10 dB higher than the total gain of the useful signal from intenna to the data output terminal. When the un-isolated signal goes though the entire communication system and gets a maximum of gain, 130 dB, it is still at a level under the sensitivity 10 dB. Then, the communication system never find its exsistence. As mentioned above, when the isolation between blocks is achieved up to 50-60 dB, it makes RFIC becomes realistic and available to applied into a system, say,

communication system. However, it is still far from the ideal goal and the research of isolation must be continued.

8.5.2 High Q Inductor for IC

The project of high Q IC inductor is the key growing point in the future RFIC as well as SOC development.

At present, the Q value of an IC inductor at RF range is about 5 to 10 in all of IC processes. It is much poorer than that of a chip or discrete inductor, in which the Q value of inductors is about 80 to 150. The low Q value of spiral inductors brings about the additional noise to all the RF blocks, and it is hard to be applied in the implementation of a filter or an oscillator.

The second problem is that a large die area must be provided to the spiral inductor. In order to reduce the cost and to simplify the hardware works, most IC inductors are constructed on the chip. The die area for each spiral inductor is about 10 to 100 times larger than the area of the device or other parts. Looking at any of existing RFIC layouts, about 60% to 80% of the area is occupied by the IC spiral inductors. The IC spiral inductor becomes the main factor of the total IC cost.

References

[1] Jacob Millman and Christos C. Halkias, “Integrated Electronics: Analog and Digital Circuits and Systems” (Book), McGraw-Hill Book Company, 1972.

[2] H. M. Greenhouse, “Design of Planar Rectangular Microelectronic Inductors,” IEEE Transaction Parts, Hybrids, and Packaging, Vol. PHP-10, No. 2, pp. 101-109, June 1974.

[3] Herbert Taub and Donald Schilling, “Digital Integrated Electronics,” (Book), McGraw-Hill, Inc., 1977.

[4] T. Sakurai, and K. Tamam, “Simple formulas for Two- and Three-Dimensional Capacitances,” IEEE Transaction Electron Devices, Vol. ED-30, No. 2, pp. 183-185, 1983.

[5] E. Barke, “Line-to-Ground Capacitance Calculation for VLSI: A Comparison,” IEEE Transaction Computer-Aided Design, Vol.7, No.2, pp.295-298, February 1988.

[6] Yun-Ti Wang, and Asad A. Abidi, “CMOS Active Filter Design at Very High Frequencies,” IEEE Journal of Solid-State Circuits, Vol.25, NO. 6, pp.1562-1574, December 1990.

[7] Paul R. Gray, Robert G. Meyer, “Analysis and Design of Analog Integrated Circuits,” (Book), Third Edition, John Wiley & Sons, Inc., 1993.

[8] Ashby, K.B.; Finley, W.C.; Bastek, J.J.; Moinian, S.; Koullias, I.A.; “70 High Q inductors for wireless applications in a complementary silicon bipolar process,” Proceedings of the 1994 Bipolar/BiCMOS Circuits and Technology Meeting,1994, pp. 179-182, October 10-11, 1994.

[9] P. Yue et al., “A Physical Model for planar Spiral Inductors on Silicon,” IEDM Proceedings, December 1996.

[10] Ashby, K.B.; Koullias, I.A.; Finley, W.C.; Bastek, J.J.; Moinian, S.; “High Q inductors for wireless applications in a complementary silicon bipolar process,” IEEE Journal of Solid-State Circuits, Vol. 31, Issue 1, pp. 4-9, January 1996.

[11] P. Yue, and S. Wong, “On-Chip Spiral Inductors with Patterned Ground Shields for Si-Based RFIC’s,” VLSI Circuits Symposium Digest of Technical Papers, June 1997.

[12] Nelson, C.; Sigliano, R.; Makihara, C.; “Built-in passive components in multilayer ceramics for wireless applications,” Proceedings of 3rd International Symposium on Advanced Packaging Materials, pp. 75-80, March 9-12, 1997.

[13] Klemmer, N.; Hartung, J.; “High Q inductors for MCM-Si technology,” 1997 IEEE Multi-Chip Module Conference, 1997. MCMC '97, pp. 33-37, February 4-5, 1997.

[14] Hirad samavati et. al., “Fractal Capacitors”, IEEE Journal of Solid-State Circuits, Vol.33, NO. 12, pp.2035-2041, December 1998.

[15] Erben, U.; Abele, P.; Behammer, D.; Ergraber, H.; Schmacher, H.; “High-Q inductors on silicon using a quasi thin film microstrip technique,” 1998 Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems, 1998. Digest of Papers., pp. 155-159, September 17-18, 1998.

[16] Sung-Mo (Steve) Kang and Yusuf Leblebici, “CMOS Digital Integrated Circuits” (Book), WCB/McGraw-Hill Companies, Inc., 1999.

[17] Groves, R.; Malinowski, J.; Volant, R.; Jadus, D.; “High Q inductors in a SiGe BiMOS process utilizing a thick metal process add-on module,” Proceedings of the 1999 Bipolar/BiCMOS Circuits and Technology Meeting, 1999., pp. 149-152, September 26-28, 1999.

[18] Frei, M.R.; Belk, N.R.; Dennis, D.C.; Carroll, M.S.; Lin, W.; Pinto, M.R.; Archer, V.D.; Ivanov, T.G.; Moinian, S.; Ng, K.K.; Chu, J.; “Integration of high-Q inductors in a latch-up resistant CMOS technology,” International Electron Devices Meeting, 1999. IEDM Technical Digest., pp. 757-760, December 5-8, 1999.

[19] Kamogawa, K.; Nishikawa, K.; Tokumitsu, T.; Tanaka, M.; “A novel high-Q inductor based on Si 3D MMIC technology and its application,” 1999 IEEE Radio Frequency Integrated Circuits (RFIC) Symposium, pp. 185-188, June 13-15. 1999.

[20] Sung-Jin Kim; Yong-Goo Lee; Sang-Ki Yun; Hai-Young Lee; “Realization of high-Q inductors using wirebonding technology,” The First IEEE Asia Pacific Conference on ASICs, 1999. AP-ASIC '99., pp. 13-16, August 23-25, 1999.

[21] U. Yodprasit, and J. Ngarmnil, “Q-Enhancing Technique for RF CMOS Active Inductor,” ISCAS 2000, IEEE International Symposium on Circuits and Systems, Geneva, May 28-31, 2000.

[22] Hannu Tenhunen, “CMOS Interconnects” (Lecture), Electronic System Laboratory, Kungl Tekniska Hogskolan, 2000.

[23] U. Yodprasit, and J. Ngarmnil, “Q-enhancing Technique for RF CMOS Active Inductor”, ISCAS 2000, Geneva Switzerland -IEEE International Symposium on Circuits and Systems, May 28-31, 2000.

[24] David, J.-B.; Musalem, F.-X.; Albert, P.; “RF high-Q spiral inductor design,” Proceedings of the First ISA/IEEE Conference, Sensor for Industry, 2001, November 5-7, 2001
Pages:78 – 82

[25] Behzad Razavi, “Design of Analog CMOS Integrated Circuits,” McGraw-Hill Companies, 2001.

[26] Jenei, S.; Decoutere, S.; Van Huylenbroeck, S.; Vanhorebeek, G.; Nauwelaers, B.; “High Q inductors and capacitors on Si substrate,” 2001 Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems, 2001. Digest of Papers. pp. 64-70, September 12-14, 2001.

[27] Theerachet Soorapanth, and S. Simon Wong, “A 0-dB IL 2140+/-30MHz Bandpass Filter Utilizing Q-enhanced Spiral Inductors in Standard CMOS”, IEEE Journal of Solid-State Circuits, Vol.37, NO. 5, pp.579-586, May 2002.

[28] Van Schuylenbergh, K.; Chua, C.L.; Fork, D.K.; Jeng-Ping Lu; Griffiths, B.; “On-chip out-of-plane high-Q inductors,” Proceedings of IEEE Lester Eastman Conference on High Performance Devices, 2002., pp. 364-373, August 6-8, 2002.

[29] Lakdawala, H.; Zhu, X.; Luo, H.; Santhanam, S.; Carley, L.R.; Fedder, G.K.; “Micromachined high-Q inductors in a 0.18-μm copper interconnect low-k dielectric CMOS process,” IEEE Journal of Solid-State Circuits, Vol. 37, Issue 3, pp. 394-403, March 2002.

[30] Rotella, F.; Howard, D.; Racanelli, M.; Zampardi, P.; “Characterizing and optimizing high Q inductors for RFIC design in silicon processes,” 2003 IEEE Radio Frequency Integrated Circuits (RFIC) Symposium, pp. 339-342, June 8-10, 2003.

Index

Contents

Chapter 9 Noise, Gain, and Sensitivity of a Receiver

9.1 Noise in a Circuit Block or a System

Noise inherently exists in either active or passive parts of an RF circuit. In the micro electrical world, the charge amount of the main electric carriers, electrons and the positive-charged holes, is not a continuous quantity and the motion of the carriers is a chaotic or random process in the time domain. The macro parameter, which describes the random process, always contains two portions. One is its average value while another is its random fluctuation from its average value. The random fluctuation can be described by its root-square-mean value or variance. For example, the electric current always contains two portions. One is the current with its average value that we are familiar with. Another one is the noise current that it is always neglected in the field where such a tiny fluctuation of current is not important. In the electronic circuit design field, the random fluctuation of the charge carriers is one of the main noise sources and should never be neglected. Noise current, noise voltage, or noise power is one of the most important parameters in the performance of a block or a system.

In a circuit or a system, another noise source comes from its outside environment. When a signal is transported or operated in a circuit block or system, it is inevitably interfered or imposed with the noise, which is radiated from other blocks or systems. Limited by the topics it will not be discussed in this book. The readers who are interested in this subject can refer to other specific books or papers.

Noise degrades the performance of the signal. One of main tasks in the circuit design is to reduce or suppress the noise added by the circuit itself or from the environment outside so as to ensure a good performance of the signal.

9.1.1 Noise Sources

There are many kinds of noise sources, such as shot noise, thermal noise, flicker noise, burst noise (popcorn noise), avalanche noise, and so on. We are going to introduce only 3 noise sources in this section.

9.1.1.1 Shot Noise

Shot noise is associated with the current itself. A current appears to be stable in the macro world whereas it appears to be fluctuated in the micro world. A current, in fact, is composed of a large number of moving carriers, either positive charges or electrons with their random velocities. The product of the electric charge and velocity of a charge or an

electron forms a current element. Owing to the different velocity, these current elements are different from each other. The sum of all the current elements at a junction of semiconductor or a cross-section of a part looks like a large number of current pulses. It is slightly different from time to time due to different number of current pulses and different velocities. The average value of these random current pulses is called current and its fluctuation around the average value is called shot noise.

Try to imagine a scene where millions and millions of soldiers are in practice of firing a gun towards approximately the same direction. The soldiers are using a variety of guns so that the bullet speeds are different. Every soldier is asked to fire the gun consecutively for a long period. The bullet flow could be imagined like a collection of the current pulses that we are concerning. The instructor, who stands at a safe place, counts the number of bullets passing over a specific cross-section. The number of bullets and their speed fluctuate around an average value. The deviation of the fluctuation is tiny but random. The current in a device, a part, or a runner, is similar to such an example.

Mathematically the product of the average drift velocity of the moving carrier and the average value of the charge passing over the unit cross section is the value of the current, and the fluctuation of current is a noise. It is named as shot noise because its fluctuation is due to a large number of random shots of moving carriers. Up to statistics, the mean-square value of its fluctuation is

$$\overline{i_n^{\,2}} = 2qI\,\Delta f \quad , \tag{9.1}$$

where i_n = Current fluctuation, a random variable,
I = Average value of current,
q = Charge of an electron =1.6 x 10^{-19} Coulomb,
Δf = Bandwidth in which noise source is acting.

The shot noise is proportional to the current itself. Shot noise exists wherever the current drain appears. The less the current drain, the lower the shot noise would be. However, the current drain is indispensable in a circuit with the active devices and very often a higher current drain is required for the better performance of the circuit. Therefore, a circuit designer must take a tradeoff between the shot noise and the high current drain.

9.1.1.2 Thermal Noise

Thermal noise is essentially the fluctuation of the resistance in the parts or devices. Electrons play the main role in the formation of resistance because they have much higher mobility than other charge carriers. The collision between electrons and between electron and other charge carriers or particles forms the resistance. The random thermal motion and collision of electrons produce the thermal noise while the random drift velocity of the moving charge-carriers or the current fluctuation produces the shot noise. The velocity of the electron thermal motion is much higher than the drift velocity of the

electron and other charge-carriers. Therefore, thermal noise is a completely different mechanism from shot noise. Thermal noise still exists even when the shot noise is gone due to a zero current.

The thermal noise of a resistor can be represented by either voltage or current noise source, that is,

$$\overline{e_n^{\,2}} = 4kTR\Delta f \quad , \tag{9.2}$$

or,

$$\overline{i_n^{\,2}} = 4kT\frac{1}{R}\Delta f \quad , \tag{9.3}$$

where e_n = Voltage fluctuation,
i_n = Current fluctuation,
k = Boltzmann constant,
T = Room Kelvin temperature,
R = Resistance of resistor,
Δf = Bandwidth in which noise source is acting.

For example, at room temperature, 300^{o}K, a 1kΩ resistor corresponds to a thermal noise spectral density 16.6 x10^{-18}*V* ^{2}/*Hz*, and a noise voltage source with its *rms* voltage value of 4 *nV*/*Hz*$^{1/2}$ or, a noise current source with the same amount as that of shot noise by a direct current 50 *uA*.

9.1.1.3 Flicker Noise (1/f Noise)

At an early stage, flicker noise is found in the emitter-base depletion layer due to contamination and crystal defects. The traps capture and release carriers in a random fashion associated a noise with energy concentrated at low frequency. Later on, it is found in all the active devices as well as in some passive parts. Its existence is always associated with a direct current. Therefore, a resistor that is not carrying a direct current does not have flicker noise. Also, a metal type of external resistor has no flicker noise, because the traps associated with contamination do not exist. It is therefore recommended if the direct current is to be carried.

Flicker noise can be represented by a noise current source as follows:

$$\overline{i_n^{\,2}} = k_f\frac{I^a}{f^b}\Delta f \quad , \tag{9.4}$$

where i_n = Current fluctuation,

k_f = A constant for a particular device,
I = Direct current,
a = 0.5 to 2,
$b \approx 1$.

Flicker noise is significant in the low frequency range. Therefore, in PLL and oscillator designs, the flicker noise is one of the main sources of phase noise.

9.1.2 Definition of Noise Figure

At the output or any point of a circuit block, the signal can be detected only if the power of signal is greater than the power of noise. The power ratio of signal to noise is a measure of possibility to detect a signal in a practical circuit.

In reality, a noise-free circuit block does not exist. A practical circuit block always produces an additional part of noise to the signal. The additional noise therefore reduces the ratio of the signal to noise down from input to output. That is, in a practical circuit block, we always have

$$\frac{S_o}{N_o} < \frac{S_i}{N_i} \quad , \tag{9.5}$$

because

$$S_o = GS_i \quad , \tag{9.6}$$

$$N_o = GN_i + \Delta N \quad , \tag{9.7}$$

then

$$\Delta N > 0 \, , \tag{9.8}$$

where ΔN = additional noise power produced by the circuit block,
S_i, S_o = signal power at input and output respectively,
N_i, N_o = noise power at input and output respectively,
G = power gain of the circuit block.

Figure 9.1 shows various parameters in input and output for a noisy block.

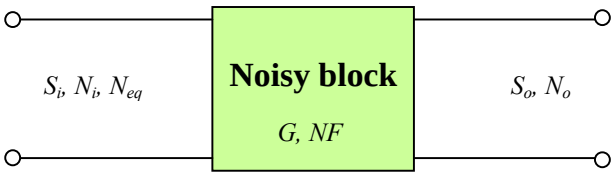


Figure 9.1 Various parameters related to the definition of noise figure.

Instead of ΔN, a new parameter, noise figure, NF (*Noise Figure*), can be defined and can also describe the additional noise produced by the circuit block, that is,

$$NF = \frac{\frac{S_i}{N_i}}{\frac{S_o}{N_o}}, \tag{9.9}$$

The noise figure, *NF*, is a good measure of the noise performance of a system or a block. It can be interpreted by other words. By the substitution of the relation (9.6) into (9.9), we have

$$NF = \frac{N_o}{GN_i} = \frac{N_o}{N_{o,i}}, \tag{9.10}$$

where $N_{o,i} = GN_i$ = Output noise power portion due to the input noise power.

Or, from expression (9.7),

$$NF = \frac{GN_i + \Delta N}{GN_i}. \tag{9.11}$$

In other words,

$$NF = \frac{\text{Total output noise power}}{\text{Output noise power due to the input noise}}.$$

9.1.3 Noise Figure in a Noisy Two Port Block

A noisy two port block can be characterized by its source noise current fluctuation and output noise current fluctuation as shown in Figure 9.2 (a). Of course, the noise current fluctuation is measured with its *rms* (root-mean-square) value since the noise current fluctuation is a random variable with its mean value equal to zero. The noise source is represented by the source noise current fluctuation i_{ns} and the source admittance Y_S. The output noise current fluctuation, i_{no}, is contributed by the source noise current fluctuation, i_{ns}, and the additional noise inside the noisy two-port block. The mean square of noise current fluctuation, either source noise current fluctuation or output noise current fluctuation, is proportional to the mean of noise power.

The additional noise inside a noisy two-port block can be characterized by an input noise voltage generator e_{ni} in series, an input noise current generator i_{ni} in parallel, and a noise-free two-port block, as shown in Figure 9.2 (b). The input noise voltage fluctuation e_{ni} in series, accounts for the output noise fluctuation i_{no} which is observed when the input is short-circuited. The second noise generator, the input noise current fluctuation i_{ni} in

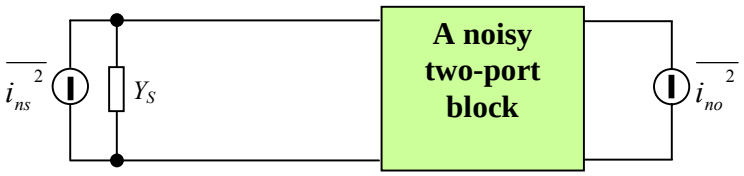
parallel, is necessary because the output noise power is always existed even when the input is open-circuited so that the input noise voltage generator e_{ni} is ignored.

The output noise now contributed by three noise generators, that is

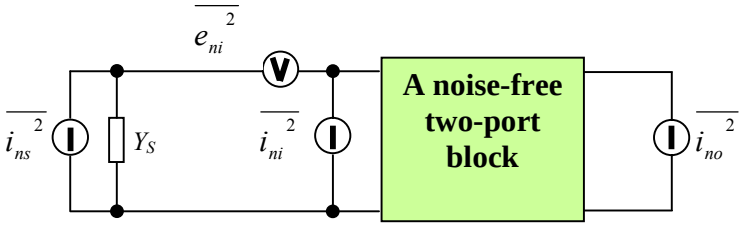
$$\overline{i_{no}^{\,2}} = G\left(\overline{i_{ns}^{\,2}} + \overline{|\, i_{ni} + Y_s e_{ni}\,|^2}\right) , \tag{9.12}$$

and $$Y_s = G_s + \mathrm{j}B_s,$$

where G_s = Conductance of source,
B_s = Susceptance of source,
G = Power gain of the noisy two-port block,



(a) Characterization of a noisy two port block by source noise current, output noise current, and a noisy two-port block.



(b) Characterization of a noisy two port block by source noise current, input noise voltage and noise current, output noise current, and a noise-free two-port block.

Figure 9.2 Noise characterization of a noisy two-port block.

The output noise due to the source noise only is

$$\left.\overline{i_{no}^{\,2}}\right|_{n.s.only} = G\left(\overline{i_{ns}^{\,2}}\right) . \tag{9.13}$$

Consequently, from the two expressions above, the noise figure of the noisy two-port block is

$$NF = \frac{\overline{i_{ns}^{\,2}} + \overline{|\, i_{ni} + Y_s e_{ni}\,|^2}}{\overline{i_{ns}^{\,2}}} = 1 + \frac{\overline{|\, i_{ni} + Y_s e_{ni}\,|^2}}{\overline{i_{ns}^{\,2}}} . \tag{9.14}$$

In the denominator, the mean square of source noise current fluctuation is related to the source conductance, G_S, by the Nyquist formula:

$$\overline{i_{ns}^{\,2}} = 4kTG_S\Delta f \quad . \tag{9.15}$$

In the numerator of expression (9.14), i_{ni} must be split into two components: One is perfectly correlated and another one is uncorrelated with the input noise voltage fluctuation, e_{ni}, that is,

$$i_{ni} = \left(i_{ni} - i_{niu}\right) + i_{niu} \quad , \tag{9.16}$$

where i_{niu} = Component of i_{ni} uncorrelated with e_{ni}, the remained component, (i_{ni}-i_{niu}), of course, is correlated with e_{ni}.

The correlated component, (i_{ni}-i_{niu}), is correlated with e_{ni} and can be expressed by a proportional constant Y_c.

$$i_{ni} - i_{niu} = Y_c e_{ni} \quad , \tag{9.17}$$

and

$$\overline{\left|i_{ni} - i_{niu}\right|^2} = Y_c^{\,2}\overline{e_{ni}^{\,2}} \quad , \tag{9.18}$$

where Y_c = correlated admittance of Y_s with e_{ni}, and is

$$Y_c = G_c + jB_c \quad , \tag{9.19}$$

where G_c = Correlated conductance of the source admittance,
B_c = Correlated susceptance of the source admittance.

Similar to expression (9.15), the mean square of uncorrelated input noise current can be expressed as

$$\overline{i_{niu}^{\,2}} = 4kTG_u\Delta f \quad . \tag{9.20}$$

where G_u = Uncorrelated conductance of the source admittance.

The input noise voltage can be expressed in terms of an equivalent noise resistance R_n,

$$\overline{e_{ni}^{\,2}} = 4kTR_n\Delta f \quad . \tag{9.21}$$

Now let's conduct a somehow tedious derivation of the numerator of expression (9.14).

$$\overline{|\,i_{ni} + Y_s e_{ni}\,|^2} = \overline{\left(i_{ni} + Y_s e_{ni}\right)\left(i_{ni}^{\,*} + Y_s^{\,*} e_{ni}^{\,*}\right)} = \overline{i_{ni}^{\,2} + Y_s^{\,2} e_{ni}^{\,2} + e_{ni} i_{ni}^{\,*} Y_s + e_{ni}^{\,*} i_{ni} Y_s^{\,*}}$$

$$\overline{|i_{ni}+Y_s e_{ni}|^2}=\overline{(i_{ni}-i_{niu})^2}+\overline{i_{niu}{}^2}+Y_s{}^2\overline{e_{ni}{}^2}+Y_c{}^*Y_s\overline{e_{ni}{}^2}+Y_cY_s{}^*\overline{e_{ni}{}^2}$$

$$\overline{|i_{ni}+Y_s e_{ni}|^2}=Y_c{}^2\overline{e_{ni}{}^2}+\overline{i_{niu}{}^2}+Y_s{}^2\overline{e_{ni}{}^2}+Y_c{}^*Y_s\overline{e_{ni}{}^2}+Y_cY_s{}^*\overline{e_{ni}{}^2}$$

$$\overline{|i_{ni}+Y_s e_{ni}|^2}=\overline{i_{niu}{}^2}+\left|Y_s+Y_c\right|^2\overline{e_{ni}{}^2}\quad . \tag{9.22}$$

In the process of the above derivation, the following relations have been applied:

$$\left(A\ +B\right)^*=A^*+B^*\ ,$$

$$AA^*=\left|A\right|^2\ ,$$

$$\left(A\ B\right)^*=A^*B^*\ ,$$

$$\overline{A\ +B}=\overline{A}+\overline{B}\ ,$$

and,

$$\overline{e_{ni}{}^*i_{ni,u}}=\overline{e_{ni}i_{niu}{}^*}=0\quad ,$$

$$\overline{e_{ni}{}^*i_{ni}}=\overline{e_{ni}{}^*\left(i_{ni}-i_{niu}\right)}=Y_c\overline{e_{ni}{}^2}\quad ,$$

$$\overline{e_{ni}i_{ni}{}^*}=\overline{e_{ni}\left(i_{ni}-i_{niu}\right)^*}=Y_c{}^*\overline{e_{ni}{}^2}\quad .$$

From (9.15) to (9.22), equation (9.14) becomes

$$NF=1+\frac{G_u}{G_s}+\frac{R_n}{G_s}\left[\left(G_s+G_c\right)^2+\left(B_s+B_c\right)^2\right]. \tag{9.23}$$

It can be seen that the noise figure is a function of the 4 parameters G_u, R_n, G_c and B_c. The first two parameters, G_u and R_n, depend upon the operating status and operating frequency of the two port block while the another two parameters, G_c and B_c, are a correlated portion of the source admittance. It is therefore more interesting to search for the minimum noise figure as the source admittance is varied. It corresponds to two minimized conditions since the source admittance consists of two portions, G_s and B_s:
From the minimized condition,

$$\frac{\partial\left(NF\right)}{\partial B_s}=0\ ,$$

by the differentiating of equation (9.23) in respect to B_s, it is found that

$$B_o = B_s = -B_c\,, \tag{9.24}$$

where B_o = Optimum value of the source susceptance, B_s, which enables *NF* to reach a minimum. Thus, equation (9.23) becomes

$$NF = 1 + \frac{G_u}{G_s} + \frac{R_n}{G_s}\left(G_s + G_r\right)^2. \tag{9.25}$$

Now, from another minimized condition,

$$\frac{\partial(NF)}{\partial G_s} = 0\,,$$

by differentiating of equation (9.25) in respect to G_s, namely

$$\frac{\partial(NF)}{\partial G_s} = \frac{G_s 2R_n\left(G_s + G_c\right) - G_u - R_n\left(G_s + G_c\right)^2}{{G_s}^2} = 0\,,$$

it is found that

$$G_o = G_s = \sqrt{\frac{G_u + R_n {G_c}^2}{R_n}}\,, \tag{9.26}$$

where G_o=Optimum value of source conductance, G_s, which also enables *NF* to reach a minimum.

Substituting of equations (9.24) and (9.26) into (9.23), the minimum of noise figure is

$$NF_{\min} = 1 + 2R_n\left(G_c + G_o\right), \tag{9.27}$$

and expression (9.23) can be re-written as

$$NF = NF_{\min} + \frac{R_n}{G_s}\left[\left(G_s - G_o\right)^2 + \left(B_s - B_o\right)^2\right]. \tag{9.28}$$

This expression of the noise figure implies a remarkable significance. It says that for any noisy two port circuit block, a minimum of noise figure is attained as long as the source admittance, $Y_s=G_s+jB_s$, is adjusted to its optimum values, $Y_o=G_o+jB_o$.

Instead of those 4 parameters, G_u, R_n, G_c, and B_c in expression (9.23), the noise figure is a function of another 4 parameters, NF_{min}, R_n, G_o, and B_o. Once these 4 parameters are known, we can calculate the noise figure for the case with any source admittance. An experimental way to obtain the 4 parameters was developed and is called the 4 points method of noise figure testing. By the means of testing noise figure 4 times with 4

different source admittances, 4 equations can be established on the basis of expression (9.28). Consequently the 4 unknown parameters, NF_{min}, R_n, G_o, and B_o can be solved.

9.1.4 Minimum Noise Figure and Equivalent Noise Resistor

In the practical engineering design, especially in an LNA design, the first approach is to reach the minimum of the noise figure. And, the second approach is to set the equivalent noise resistor as low as possible so that the a low value of the final noise figure can be obtained. As an example, we would discuss the cases when the noisy two port block is built by a MOSFET or a bipolar device.

9.1.4.1 Noise in a MOSFET

There are two noise sources in a MOSFET: the drain noise current and the gate noise current. Their mean-square values can be expressed by

$$\overline{i_{nd}^{\,2}} = 4kT\gamma g_{do}\Delta f \,, \tag{9.29}$$

$$\overline{i_{ng}^{\,2}} = 4kT\delta g_{g}\Delta f \,, \tag{9.30}$$

respectively,

where i_{nd} = Drain current noise fluctuation, a random variable,

i_{ng} = Gate current noise fluctuation, a random variable,

γ = 1 for short channel device at V_{DS}=0 and decreases toward 2/3 for long channel device in saturation ,

δ = Coefficient to be about twice of γ,

g_{do} = Drain-source conductance at zero V_{DS},

g_g = Gate-source conductance and is

$$g_g = \frac{\omega^2 C_{gs}^{\,2}}{5g_{do}}. \tag{9.31}$$

The study of noise in a MOSFET is to find four noise parameters, NF_{min}, R_n, G_o, and B_o in the equation (9.28). They have been derived and formalized in many papers and books. We would like to recommend readers to read the corresponding sections in Thomas's book (Thomas H. Lee, The Design of CMOS Radio-Frequency Integrated Circuit, Cambridge University Press, 1998). The expressions of these four noise parameters are

The equivalent noise resistance R_n,

$$R_n = \frac{\gamma g_{do}}{g_m^{\,2}} \quad . \tag{9.32}$$

The optimum source conductance G_o,

$$G_o = \frac{g_m}{g_{do}} \omega C_{gs} \sqrt{\frac{\delta}{5\gamma}\left(1-|c|^2\right)} \quad . \tag{9.33}$$

The optimum source susceptance B_o,

$$B_o = -\omega C_{gs}\left(1+\frac{g_m}{g_{do}}|c|\sqrt{\frac{\delta}{5\gamma}}\right) \quad . \tag{9.34}$$

Finally, the minimum of noise figure NF_{min},

$$NF_{\min} = 1+\frac{2}{\sqrt{5}}\frac{\omega}{\omega_T}\sqrt{\gamma\delta\left(1-|c|^2\right)} \quad , \tag{9.35}$$

where

c is a correlation coefficient between the gate noise and the drain noise , that is

$$c = \frac{\overline{i_{ng}.i_{nd}^{\,*}}}{\sqrt{\overline{i_{ng}^{\,2}}.\overline{i_{nd}^{\,2}}}} \quad . \tag{9.36}$$

The coefficient c is about j0.395 for a long channel device.

g_m is defined as

$$\overline{e_{ni}^{\,2}} = \frac{\overline{i_{nd}^{\,2}}}{g_m^{\,2}} = \frac{4kT\gamma g_{do}\Delta f}{g_m^{\,2}} \quad , \tag{9.37}$$

and ω_T is approximately equal to the ratio of gm and Cgs, that is

$$\omega_T \approx \frac{g_m}{C_{gs}} \quad . \tag{9.38}$$

9.1.4.2 Noise in a Bipolar Device

The study of noise in a bipolar transistor has been derived and formalized in many papers and books. We would like to recommend readers to read the corresponding sections in Paul's book (Paul R. Gray, Paul J. Hurst, Stephen H. Lewis, and Robert G. Meyer,

"Analysis and Design of Analog Integrated Circuits," (Book), Fourth Edition, John Wiley & Sons, Inc., 2001.). The expressions of these four noise parameters are

Figure 9.3 illustrates the noise figure contributed by the source and the equivalent internal noise of a bipolar device.

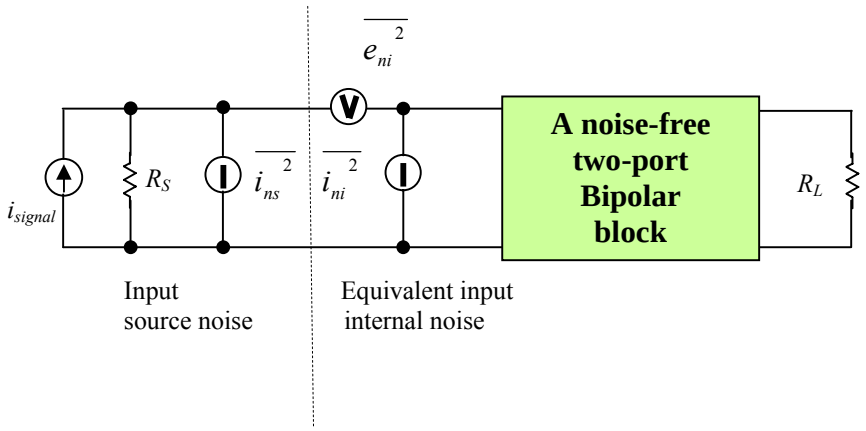


Figure 9.3 Noise figure contributed by source noise and equivalent internal noise

It is assumed that
1) The source impedance is a pure resistor R_s,
2) Only thermal noise and shot noise are considered while the flicker noise and other high frequency effects are neglected.

The noise sources are
1) Source noise

$$\overline{i_{ns}^2} = 4kT\frac{1}{R_s}\Delta f \quad , \tag{9.39}$$

2) Equivalent internal voltage noise

$$\overline{v_{ni}^2} = 4kT\left(r_b + \frac{1}{2g_m}\right)\Delta f \quad , \tag{9.40}$$

3) Equivalent internal current noise

$$\overline{i_{ni}^2} = 2qI_B\Delta f = 2q\frac{I_C}{\beta_F}\Delta f \quad , \tag{9.41}$$

It is found that there is a minimum of NF, NF_{min}, when the source resistance is

$$R_s = R_{s,opt} = \frac{\sqrt{\beta_F}}{g_m}\sqrt{1+2g_m r_b} \quad . \tag{9.42}$$

then,

$$NF_{\min} \approx 1 + \frac{1}{\sqrt{\beta_F}} \sqrt{1 + 2 g_m r_b} \quad . \tag{9.43}$$

There are 3 parameters, β_F, r_b, and g_m, which determine the value of NF_{min}. The higher the β_F, the lower the NF_{min}. The lower r_b, the lower the NF_{min}. The lower g_m, the lower the NF_{min} because g_m is proportional to I_c which brings about the shot noise.

9.2 Gain

Generally there are 3 well-known gains in an electronic circuit block: power gain, voltage gain, and current gain. The meaning of the voltage and the current gain is quite straight forward. It is simply the ratio between input and output. However, there are different power gains with different meanings.

9.2.1 Definition of Power Gains

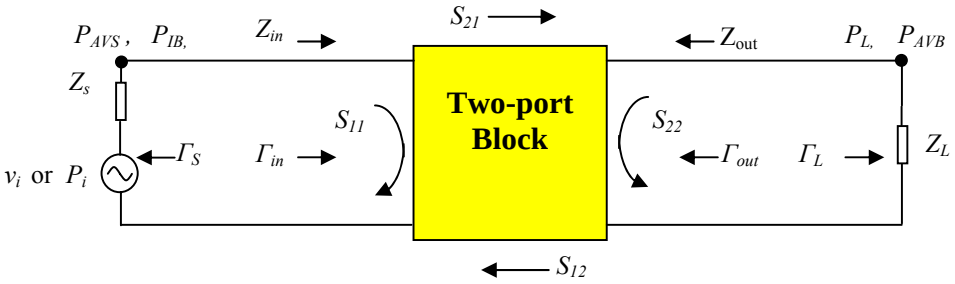


Figure 9.4 Relationship between power gain and related parameters

Figure 9.4 shows a two-port block in a general case when

$$\Gamma_s \neq 0 \quad , \tag{9.44}$$

and

$$\Gamma_L \neq 0 \quad , \tag{9.45}$$

It is well-known that

$$\Gamma_{in} = S_{11} + \frac{S_{12}S_{21}\Gamma_L}{1 - S_{22}\Gamma_L} = \frac{Z_{in} - Z_o}{Z_{in} + Z_o} \quad , \tag{9.46}$$

$$\Gamma_{out} = S_{22} + \frac{S_{12}S_{21}\Gamma_s}{1 - S_{11}\Gamma_s} = \frac{Z_{out} - Z_o}{Z_{out} + Z_o} \quad , \tag{9.47}$$

where Z_o = Reference impedance when the S parameters are measured and is usually 50 Ω.
Γ_s = Voltage reflection coefficient at source,
Γ_L = Voltage reflection coefficient at load,
Γ_{in} = Voltage reflection coefficient at input of two-port block,
Γ_{out} = Voltage reflection coefficient at output of two-port block.

A signal flow graph as shown in Figure 9.5 is a powerful tool in analysis of the transmission and reflection of the power. Each variable is assigned as a node. All independent variables or nodes are denoted by the symbol a_i, such as a_1, a_2, a_3… All of dependent variables or nodes are denoted by the symbol b_i, such as b_1, b_2, b_3……. The relationship between the variables can be obtained by the Manson rule.

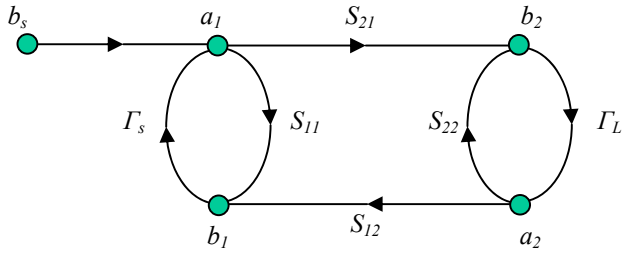


Figure 9.5 Signal flow graph of a two-port block

The powers at different nodes can be found. They are:

- Power delivered to the load P_L ,
- Power input to the block P_{IB} ,
- Power available from the source, P_{AVS} , and
- Power available from the block, P_{AVB} ,

$$P_L = \frac{|b_s|^2}{|1-\Gamma_{in}\Gamma_S|^2}|S_{21}|^2\frac{1-|\Gamma_L|^2}{|1-S_{22}\Gamma_L|^2} \quad , \tag{9.48}$$

or,

$$P_L = \frac{|b_s|^2}{|1-S_{11}\Gamma_S|^2}|S_{21}|^2\frac{1-|\Gamma_L|^2}{|1-\Gamma_{out}\Gamma_L|^2} \quad , \tag{9.49}$$

because

$$\frac{|1-S_{11}\Gamma_S|^2}{|1-S_{22}\Gamma_L|^2} = \frac{|1-\Gamma_{in}\Gamma_S|^2}{|1-\Gamma_{out}\Gamma_L|^2} \quad . \tag{9.50}$$

$$P_{AVB} = P_L\Big|_{\Gamma_L=\Gamma_{out}^*} = \frac{|b_s|^2}{|1-S_{11}\Gamma_S|^2}|S_{21}|^2\frac{1}{1-|\Gamma_{out}|^2} \quad , \tag{9.51}$$

$$P_{IB} = \frac{|b_s|^2}{|1-\Gamma_{in}\Gamma_s|^2} \quad , \tag{9.52}$$

$$P_{AVS} = P_{IB}\Big|_{\Gamma_{in}=\Gamma_S^*} = \frac{|b_s|^2}{1-|\Gamma_s|^2} \quad , \tag{9.53}$$

The transducer power gain is defined as a ratio of the power delivered to the load, P_L, to the power available from the source P_{AVS} , that is,

$$G_T = \frac{P_L}{P_{AVS}} = \frac{1-|\Gamma_s|^2}{|1-\Gamma_{in}\Gamma_s|^2}|S_{21}|^2\frac{1-|\Gamma_L|^2}{|1-S_{22}\Gamma_L|^2} \quad , \tag{9.54}$$

or,

$$G_T = \frac{P_L}{P_{AVS}} = \frac{1-|\Gamma_s|^2}{|1-S_{11}\Gamma_s|^2}|S_{21}|^2\frac{1-|\Gamma_L|^2}{|1-\Gamma_{out}\Gamma_L|^2} \quad , \tag{9.55}$$

The operating power gain is defined as a ratio of the power delivered to the load P_L to the power input to the block, P_{IB} , that is,

$$G_P = \frac{P_L}{P_{IB}} = \frac{1}{1-|\Gamma_{in}|^2}|S_{21}|^2\frac{1-|\Gamma_L|^2}{|1-S_{22}\Gamma_L|^2} \quad . \tag{9.56}$$

The available power gain is defined as the ratio of the power available from the block, P_{AVB}, to the power available from the source, P_{AVS} , that is,

$$G_A = \frac{P_{AVB}}{P_{AVS}} = \frac{1-|\Gamma_s|^2}{|1-S_{11}\Gamma_s|^2}|S_{21}|^2\frac{1}{1-|\Gamma_{out}|^2} \quad . \tag{9.57}$$

In the special case when the block becomes unilateral, that is,
when

$$S_{12} = 0 \quad , \tag{9.58}$$

then,

$$\Gamma_{in} = S_{11} \quad , \tag{9.59}$$

$$\Gamma_{out} = S_{22} \quad , \tag{9.60}$$

$$G_T = G_{T,S_{12}=0} = G_S G_o G_L = \frac{1-|\Gamma_s|^2}{|1-S_{11}\Gamma_s|^2}|S_{21}|^2\frac{1-|\Gamma_L|^2}{|1-S_{22}\Gamma_L|^2} \quad , \tag{9.61}$$

where

$$G_S = \frac{1-|\Gamma_s|^2}{|1-S_{11}\Gamma_s|^2} \quad , \tag{9.62}$$

$$G_o = |S_{21}|^2 \quad , \tag{9.63}$$

$$G_L = \frac{1-|\Gamma_L|^2}{|1-S_{22}\Gamma_L|^2} \quad . \tag{9.64}$$

It can be seen that the transducer power gain G_T is composed of 3 different independent power gains when $S_{12}=0$. And,

$$G_P = G_{P,S_{12}=0} = \frac{1}{1-|S_{11}|^2}|S_{21}|^2 \frac{1-|\Gamma_L|^2}{|1-S_{22}\Gamma_L|^2} \quad . \tag{9.65}$$

$$G_A = G_{A,S_{12}=0} = \frac{1-|\Gamma_s|^2}{|1-S_{11}\Gamma_s|^2}|S_{21}|^2 \frac{1}{1-|S_{22}|^2} \quad . \tag{9.66}$$

Furthermore, let's go to an ideal case as shown in Figure 9.6, in which the block is not only unilateral but also reflection-less at both the source and the load, that is

$$S_{12} = 0 \quad , \tag{9.67}$$

and

$$\Gamma_s = \Gamma_L = 0 \quad , \tag{9.68}$$

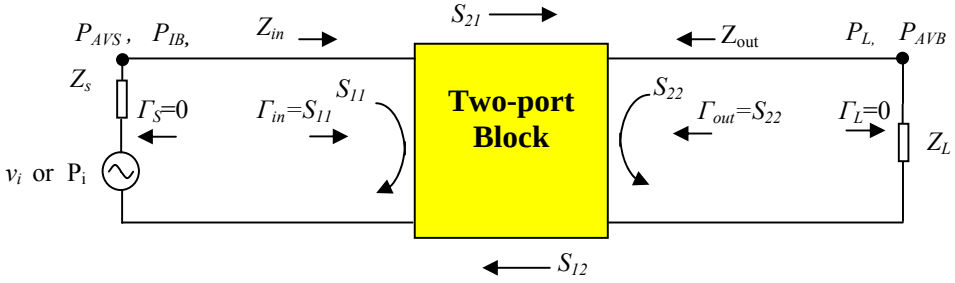


Figure 9.6 Relationship between power gain and related parameters

then,

$$G_T = G_{T,\Gamma_S=\Gamma_L=0} = G_o = |S_{21}|^2 \quad , \tag{9.69}$$

$$G_S = G_{S,\Gamma_S=\Gamma_L=0} = 1 \quad , \tag{9.70}$$

$$G_L = G_{L,\Gamma_S=\Gamma_L=0} = 1 \quad , \tag{9.71}$$

$$G_P = G_{P,\Gamma_S=\Gamma_L=0} = \frac{1}{1-|S_{11}|^2}|S_{21}|^2 \quad . \tag{9.72}$$

$$G_A = G_{A,\Gamma_S=\Gamma_L=0} = \frac{1}{1-|S_{22}|^2}|S_{21}|^2 \quad . \tag{9.73}$$

The design engineers usually present their simulation result of $|S_{21}|^2$ as power gain. It implies that it is a transducer power gain G_o, either in the unilateral case, S_{12}=0, or in the case without reflection at both of the source and the load, $\Gamma_S = \Gamma_L$=0.

9.2.2 Power Gain and Voltage Gain

In a digital circuit design, the main task is to realize the digital status transportation. The engineers always try to reduce power consumption as much as possible. Power gain is not important in the digital circuit design. Instead, the voltage gain is one of the important parameters, by which the status transportation is executed. On the other hand, in an RF/RFIC circuit design, the main task is to realize the power transportation. Consequently, it is always concerned about power gain. The voltage gain is meaningless until the corresponding impedance is specified. As long as the impedance is given, one can transfer the voltage gain to power gain through the well-known relation:

$$G_P = \frac{\frac{V_{out}^{\ 2}}{Z_{out}}}{\frac{V_{in}^{\ 2}}{Z_{in}}} = G_v^{\ 2}\frac{Z_{in}}{Z_{out}} \quad . \tag{9.74}$$

where G_p and G_v is power and voltage gain respectively.

9.3 Sensitivity

The receiver sensitivity of a communication system is mainly determined by the noise figure, the gain, and the bandwidth of the receiver.

9.3.1 Standard Noise Source

In the definition of the noise figure in the section 9.1 or in the testing of a receiver's sensitivity, a noise power source is applied, that is,

$$N_i = kT\Delta f \,, \tag{9.75}$$

where N_i = Input noise power,
k = Boltzmann's constant,
T = Absolute temperature,
Δf = Operating bandwidth of the resistor.

The theoretical background is that an actual noise source is made of a pure resistor. It is well-known that a resistor without current flowing on has thermal noise. Its mean-square of noise voltage fluctuation is

$$\overline{e_{nt}^{\,2}} = 4kTR\Delta f \,, \tag{9.76}$$

where e_{nt} = Thermal noise voltage fluctuation of the resistor,
R = The value of the resistor,

From the expression (9.76), the input noise power matched from this noise source to the input of receiver, N_i, is

$$N_i = \frac{\overline{e_{nt}^{\,2}}}{4R} = kT\Delta f \,, \tag{9.77}$$

which results the noise power source as shown in (9.75).

Figure 9.7 shows the noise source and the noise power appearing at the input of the block or receiver. It should be noted that N_i depends on the bandwidth but not a given frequency so that it is a white noise.

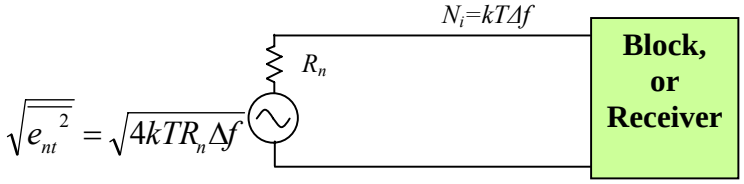


Figure 9.7 Noise source and noise power in the testing of a receiver or a block.

9.3.1 Equivalent Input Noise

We will now define another parameter N_{eq} the equivalent input noise power of the output noise power,

$$N_{eq} = \frac{N_o}{G} \quad , \tag{9.78}$$

By means of equation (9.78), expression (9.6) becomes

$$NF = \frac{N_o}{GN_i} = \frac{N_{eq}}{N_i} = \frac{N_{eq}}{kT\Delta f} \quad . \tag{9.79}$$

The noise figure NF now becomes the ratio of two input noise powers, N_{eq} and N_i. The value of N_{eq} is generally greater than $N_i = kT\Delta f$ because the additional noise power of the receiver, ΔN, is added to the output noise power N_o as shown in expression (9.7).

9.3.2 Sensitivity of a Receiver

For a given output signal to noise ratio, the minimum of a detectable input signal level is referred to as the sensitivity of the receiver. The input signal level can be measured by the power with units of dB_m, or by the voltage with unit of μV under the specified input impedance. Let's outline the sensitivity measurement in the actual engineering field.

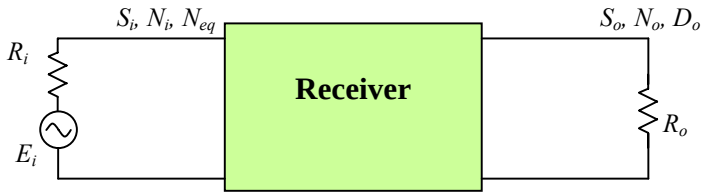


Figure 9.8 The basic parameters in the sensitivity testing for a receiver

In the output power testing, the equipment, such as the distortion meter, is able to separate the signal power from the composed power of the distortion and the noise. It is therefore an appropriate parameter being defined as *SINAD*, which is a ratio of the output signal power to the output composed power of noise and distortion, that is,

$$SINAD = \frac{S_o}{N_o + D_o} . \tag{9.80}$$

In the input portion, another ratio, called *RISE*, the ratio of the sum of input signal power and the input equivalent noise power, (S_i+N_{eq}), to the input equivalent noise power, N_{eq}, that is,

$$RISE = \frac{S_i + N_{eq}}{N_{eq}}. \tag{9.81}$$

It indicates how strong the input signal power over the input equivalent noise power.

Note that N_{eq} has been defined in equation (9.78), and the input signal power can be expressed by the parameters as shown in Figure 9.8.

$$S_i = \frac{E_i^{\,2}}{4R_i}, \tag{9.82}$$

From (9.81), we have

$$\frac{S_i}{N_{eq}} = RISE - 1. \tag{9.83}$$

Then, from (9.82) and (9.83), we have

$$E_i = 2\sqrt{S_i R_i} = 2\sqrt{N_{eq}(RISE-1)R_i},$$

$$E_i = 2\sqrt{NF(RISE-1)R_i kT\Delta f}, \tag{9.84}$$

or,

$$S_i = \frac{E_i^{\,2}}{4R_i} = NF(RISE-1)kT\Delta f. \tag{9.85}$$

where S_i, S_o = Signal power at input and output respectively, dB_w or dB_m;
N_i,N_o = Noise power at input and output respectively, *dB*;
D_o = Distortion power at output, *dB*;
R_i = Iutput resistance of generator, Ω;
N_{eq} = Equivalent input noise power of output noise;
E_i = Sensitivity looking from input, μV.

The sensitivity, E_i or S_i , is typically obtained by the measurement when the output of the receiver reaches 12 *dB SINAD*. Usually the sensitivity E_i is expressed in μV and S_i is expressed in dB_m. As shown in equations (9.84) and (9.85), the sensitivity is not only related to the noise figure but also to the bandwidth of the receiver. In general, the sensitivity in a wideband receiver is lower than that in a narrow band receiver.

References

[1] H. A. Haus et. al., "Representation of Noise in Linear Twoports," Proceedings of the IRE, Vol. 48, pp.69-74, January, 1960.

[2] Jacob Millman and Christos C. Halkias, "Integrated Electronics: Analog and Digital Circuits and Systems," (Book), McGraw-Hill Book Company, 1972.

.

[3] Ralph S. Carson, "High-Frequency Amplifiers," (Book), John Wiley & Sons, Inc., 1975.

[4] R. E. Ziemer and W. H. Tranter, "Principles of Communications," (Book), Houghton Mifflin Company, 1976.

[5] Herbert Taub and Donald Schilling, "Digital Integrated Electronics," (Book), McGraw-Hill, Inc., 1977.

[6] B. P. Lathi, "The Design of CMOS Radio-Frequency Integrated Circuits," (Book), Cambridge University Press, 1983.

[7] G. Gonzales, "Microwave Transistor Amplifier Analysis and Design," (Book), Prentice-Hall, 1984.

[8] Jack Smith, "Modern Communication Circuits," (Book), McGraw-Hill Publishing Company, 1986

[9] Lawrence E. Larson, "Modern Digital and Analog Communication Systems," (Book), Second Edition, Holt, Rinehart and Winston, Inc., 1989.

[10] D. E. Meer, "Noise figures [two-port linear transducers]," Aerospace and Electronics Conference, 1989. NAECON 1989., Proceedings of the IEEE 1989 National , Vol. 1, pp. 269 – 276, 22-26 May 1989.

[11] George D. Vendelin, Anthony M. Pavio, and Ulrich L. Rohde, "Microwave Circuit Design Using Linear and Nonlinear Techniques," (Book), John Wiley & Sons, Inc., 1990.

[13] C. E. Collins, R. Dildine, R. D. Pollard, and R. E. Miles, " A new method for determination of single sideband noise figure," Microwave Symposium Digest, 1994., IEEE MTT-S International , Vol. 3, pp. 1357 – 1360, 23-27 May 1994.

[14] Thomas H. Lee, "The Design of CMOS Radio-Frequency Integrated Circuits," (Book), Cambridge University Press, 1998.

[15] Sung-Mo (Steve) Kang and Yusuf Leblebici, "CMOS Digital Integrated Circuits," (Book), WCB/McGraw-Hill Companies, Inc., 1999.

[16] A. Geens, and Y. Rolain, "Noise figure measurements on nonlinear devices," Instrumentation and Measurement, IEEE Transactions on , Vol. 50 , Issue: 4 , pp. 971 – 975, August 2001.

[12] Paul R. Gray, Paul J. Hurst, Stephen H. Lewis, and Robert G. Meyer, "Analysis and Design of Analog Integrated Circuits," (Book), Fourth Edition, John Wiley & Sons, Inc., 2001.

[17] Phillip, E. Allen and Douglas R. Holberg, "CMOS Analog Circuit Design," (Book), Second Edition, Oxford University Press, Inc., 2002.

Index

Contents

Chapter 10 Non-linearity and Spurious Products

10.1 Spurious Products

In reality, any linear device or system is just a reasonable approximation. Strictly speaking, an ideal linear device or system does not exist. The non-linearity exists more or less in a practical device or a system. As long as the non-linearity of the devices or system exists, the spurious products are produced though in most cases they are not welcome. There are two kinds of spurious products: harmonics and complicated spurious products.

10.1.1 Harmonics

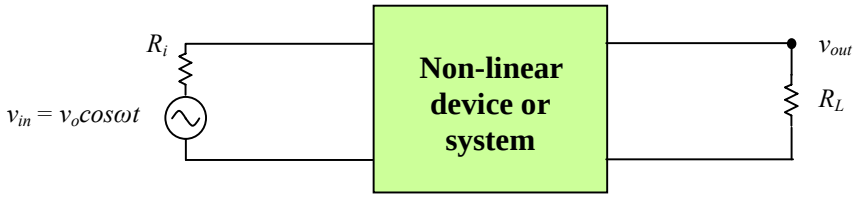


Figure 10.1 The spurious products are produced from the non-linearity in the device or system when the input is a signal with only one frequency.

In general, the non-linearity of a device or a system can be characterized by the relationship between its input and output. If the input voltage of signal is a pure sinusoidal as shown in Figure 10.1, that is,

$$v_{in} = v_o \cos \omega t \quad , \tag{10.1}$$

then the output voltage is a complicated entity with many harmonics and can be expressed by a power series of its input with corresponding coefficients as follows:

$$\begin{aligned}
v_{out} &= v_{dc} + a_1 v_{in} + a_2 v_{in}^2 + a_3 v_{in}^3 + a_4 v_{in}^4 + \ldots \\
&= V_{dc} + a_1 v_o cos\omega t + a_2 v_o^2 cos^2 \omega t + a_3 v_o^3 cos^3 \omega t + a_4 v_o^4 cos^4 \omega t + \ldots \\
&= V_{dc} + \\
&\quad + a_1 v_o cos\omega t + \\
&\quad + a_2 v_o^2 [1/2 + (1/2) cos2\omega t] + \\
&\quad + a_3 v_o^3 [(3/4) cos\omega t + (1/4) cos3\omega t] + \\
&\quad + a_4 v_o^4 [3/8 + (1/2) cos2\omega t + (1/8) cos4\omega t] + \\
&\quad + \ldots\ldots \\
&= [V_{dc} + (1/2) a_2 v_o^2 + (3/8) a_4 v_o^4 + \ldots] + \\
&\quad + [a_1 v_o + (3/4) a_3 v_o^3 + \ldots] cos\omega t + \\
&\quad + [(1/2) a_2 v_o^2 + (1/2) a_4 v_o^4 + \ldots] cos2\omega t + \\
&\quad + [(1/4) a_3 v_o^3 + \ldots] cos3\omega t + \\
&\quad + [(1/8) a_4 v_o^4 + \ldots] cos4\omega t + \\
&\quad + \ldots
\end{aligned} \tag{10.2}$$

where V_{dc}, a_1, a_2, a_3… are the coefficients of the corresponding harmonics respectively. These coefficients actually describe the non-linearity of the device or the system. They may be dependent or independent of the current flowing through the device or system, or, of the voltage drop across the device or system. By a general trend, they are reduced as its order is increased. However, sometimes there is an extra-ordinary case, in which the non-linearity coefficient with higher order has higher value than that with lower order.

From equation (10.2) it can be seen that the infinite number of harmonics, the terms containing of factor $\cos\omega t$, $\cos 2\omega t$, $\cos 3\omega t$, $\cos 4\omega t$, etc., are produced due to the non-linearity of the device or system. The new resulting coefficient of each harmonic, including DC and 1st harmonic, is another infinitive series containing the original non-linearity coefficients. These new coefficients look very clumsy at the first glance. In a practical engineering design, it is good enough to keep the first 2 or 3 terms containing the old coefficient with low orders. From equation (10.2) one can find out a remarkable feature of these new coefficients, that is, the coefficients of the odd harmonics contain only old coefficients with odd orders while the coefficients of the even harmonics contain only old coefficients with even orders. This is important to the engineers who are working for the cancellation of spurious products.

In the RF or RFIC circuit, such as LNA or PA, where a high linearity is required, the device with higher a_1 but less other coefficients should be chosen. In the circuit design of mixer, which operates on the basis of the second order of non-linearity, the device with higher a_2 but less other coefficients should be chosen.

In an actual spectrum testing for a device or a system, as long as the noise floor of the spectrum analyzer is low enough, the harmonics can be seen on the screen as shown in Figure 10.2 since non-linearity always exists.

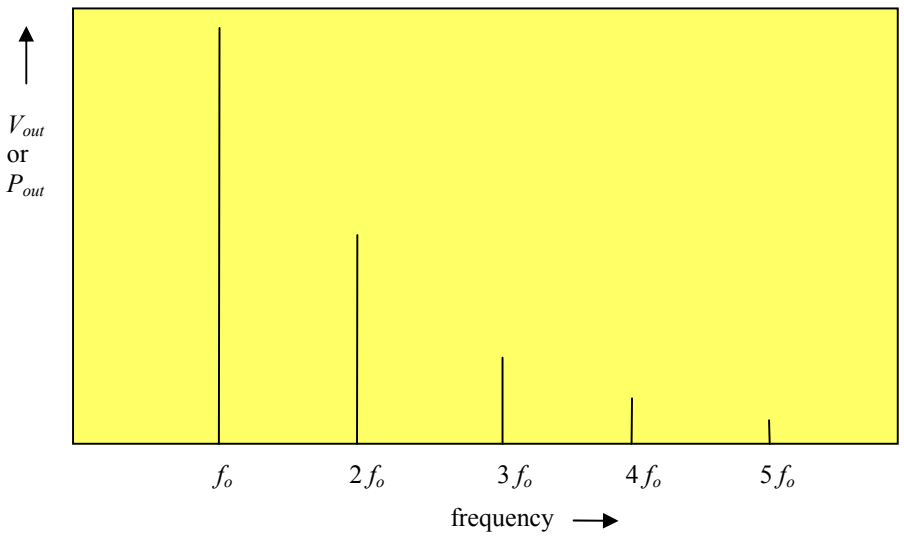


Figure 10.2 The frequency spectrum with harmonics when the input is a pure single frequency.

10.1.2 Complicated Spurious Products

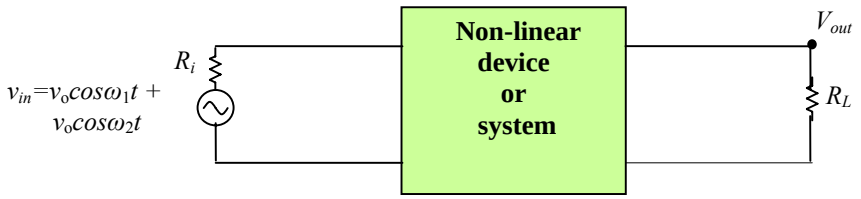


Figure 10.3 Spurious products are produced from the non-linearity in a system when the input is a signal with two frequencies

If the input voltage signal is a signal with two frequencies as shown in Figure 10.3, that is,

$$v_{in} = v_o \cos\omega_1 t + v_o \cos\omega_2 t \quad , \qquad (10.3)$$

then the output voltage is a complicated transfer function with many harmonics and composed frequencies. It can be derived as follows:

$$\begin{aligned}
V_{out} &= V_{dc} + a_1 V_{in} + a_2 V_{in}^2 + a_3 V_{in}^3 + a_4 V_{in}^4 + \ldots \\
&= V_{dc} + a_1 v_o [cos\omega_1 t + cos\omega_2 t] + a_2 v_o^2 [cos\omega_1 t + cos\omega_2 t]^2 + a_3 v_o^3 [cos\omega_1 t + cos\omega_2 t]^3 \\
&\quad + a_4 V_o^4 [cos\omega_1 t + cos\omega_2 t]^4 + \ldots\ldots \\
&= V_{dc} + a_1 v_o [cos\omega_1 t + cos\omega_2 t] + \\
&\quad + a_2 v_o^2 [1 + (1/2)cos2\omega_1 t + (1/2)cos2\omega_2 t + cos(\omega_1+\omega_2)t + cos(\omega_1-\omega_2)t] + \\
&\quad + a_3 v_o^3 [(9/4)cos\omega_1 t + (9/4)cos\omega_2 t + (1/4)cos3\omega_1 t + (1/4)cos3\omega_2 t + (3/4)cos(2\omega_1+\omega_2)t + \\
&\qquad + (3/4)cos(2\omega_2+\omega_1)t + (3/4)cos(2\omega_1-\omega_2)t + (3/4)cos(2\omega_2-\omega_1)t] + \\
&\quad + a_4 v_o^4 [9/4 + 2cos2\omega_1 t + 2cos2\omega_2 t + (1/8)cos4\omega_1 t + (1/8)cos4\omega_2 t + 3cos(\omega_1+\omega_2)t + \\
&\qquad + 3cos(\omega_1-\omega_2)t + (1/2)cos(3\omega_1+\omega_2)t + (1/2)cos(\omega_1+3\omega_2)t + (1/2)cos(3\omega_1-\omega_2)t + \\
&\qquad + (1/2)cos(3\omega_2-\omega_1)t + (3/4)cos2(\omega_1+\omega_2)t + (3/4)cos2(\omega_1-\omega_2)t] + \\
&\quad + \ldots \\
&= [V_{dc} + a_2 V_o^2 + (9/4)\, a_4 V_o^4] + \\
&\quad + [a_1 v_o + (9/4)\, a_3 v_o^3 + \ldots]cos\omega_1 t + [a_1 v_o + (9/4)\, a_3 v_o^3 + \ldots]cos\omega_2 t + \\
&\quad + [(1/2)\, a_2 v_o^2 + 2a_4 v_o^4 + \ldots]cos2\omega_1 t + [(1/2)\, a_2 v_o^2 + 2a_4 v_o^4 + \ldots]cos2\omega_2 t + \\
&\quad + [a_2 v_o^2 + 3a_4 v_o^4 + \ldots]cos(\omega_1-\omega_2)t + [a_2 v_o^2 + 3a_4 v_o^4 + \ldots]cos(\omega_1+\omega_2)t + \\
&\quad + [a_3 v_o^3 + \ldots]cos3\omega_1 t + [a_3 v_o^3 + \ldots]cos3\omega_2 t + \\
&\quad + [(3/4)\, a_3 v_o^3 + \ldots]cos(2\omega_2-\omega_1)t + [(3/4)\, a_3 v_o^3 + \ldots]cos(2\omega_2+\omega_1)t + \\
&\quad + [(3/4)\, a_3 v_o^3 + \ldots]cos(2\omega_1-\omega_2)t + [(3/4)\, a_3 v_o^3 + \ldots]cos(2\omega_1+\omega_2)t + \\
&\quad + [(1/8)\, a_4 v_o^4 + \ldots]cos4\omega_1 t + [(1/8)\, a_4 v_o^4 + \ldots]cos4\omega_2 t + \\
&\quad + [(1/2)\, a_4 v_o^4 + \ldots]cos(3\omega_2-\omega_1)t + [(1/2) a_4 v_o^4 + \ldots]cos(3\omega_2+\omega_1)t + \\
&\quad + [(1/2)\, a_4 v_o^4 + \ldots]cos(3\omega_1-\omega_2)t + [(1/2)\, a_4 v_o^4 + \ldots]cos(3\omega_1+\omega_2)t + \\
&\quad + [(3/4)\, a_4 v_o^4 + \ldots]cos2(\omega_1-\omega_2)t + [(3/4)\, a_4 v_o^4 + \ldots]cos2(\omega_1+\omega_2)t + \\
&\quad + \ldots
\end{aligned}$$

(10.4)

The resulting output contains not only the harmonics of the two frequencies but also other combined spurious products. The order of the spurious products is the number of frequencies involved in the terms, which can be listed in the Table 10.1.

Table 10.1 Order of the spurious products and their frequencies

Frequencies	Order of the spurious products
ω_1	1
ω_2	1
$2\,\omega_1$	2
$2\,\omega_2$	2
$\omega_1 - \omega_2$	2
$3\omega_1$	3
$3\omega_2$	3
$2\,\omega_1 - \omega_2$	3
$2\,\omega_1 + \omega_2$	3
$2\,\omega_2 - \omega_1$	3
$2\,\omega_2 + \omega_1$	3
$4\omega_1$	4
$4\omega_2$	4
$3\omega_2 - \omega_1$	4
$3\omega_2 + \omega_1$	4
$3\omega_1 - \omega_2$	4
$3\omega_1 + \omega_2$	4
$2(\omega_1 - \omega_2)$	4
$2(\omega_1 + \omega_2)$	4

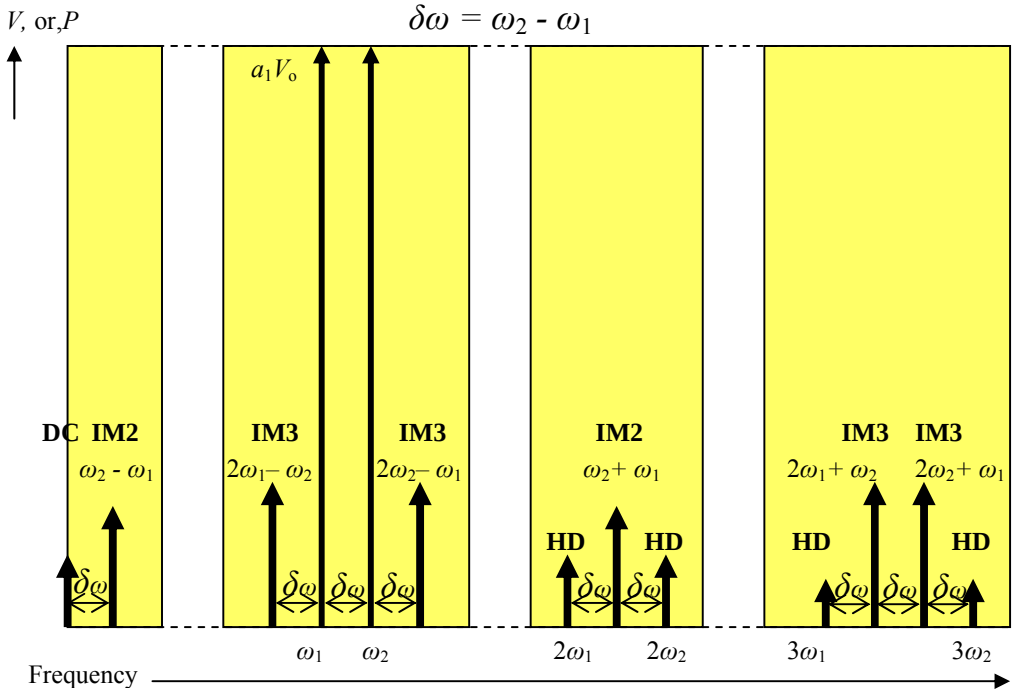


Figure 10.4 The frequency spectrum with harmonics when input is a signal with two frequencies.

Again, from equation (10.4), one can find out a remarkable feature of these new coefficients, that is, the coefficients of the odd spurious products or harmonics contain only old coefficients with odd orders while the coefficients of the even spurious products or harmonics contain only old coefficients with even orders.

Either from equation (10.4) or from Figure 10.4, it can be seen that the spectrum becomes quite complicated. The frequencies of the spurious products can be summarized as

$$\omega_{spur} = \left| m\,\omega_1 - n\,\omega_2 \right| \quad , \tag{10.5}$$

where m, n = zero and positive integers. However, it is meaningless if both of m or n are equal to 0. One of m or n must be non-zero and the spurious products are harmonics. In the case when m and n are positive integers, the spurious products are the products with composed frequencies. For example, in an LNA design, the 3rd order inter-modulation is due to the spurious product with

$$m = 2 \quad , \quad \text{and} \quad n = 1 \quad , \tag{10.6}$$

or,

$$m = 1 \quad , \quad \text{and} \quad n = 2 \quad , \tag{10.7}$$

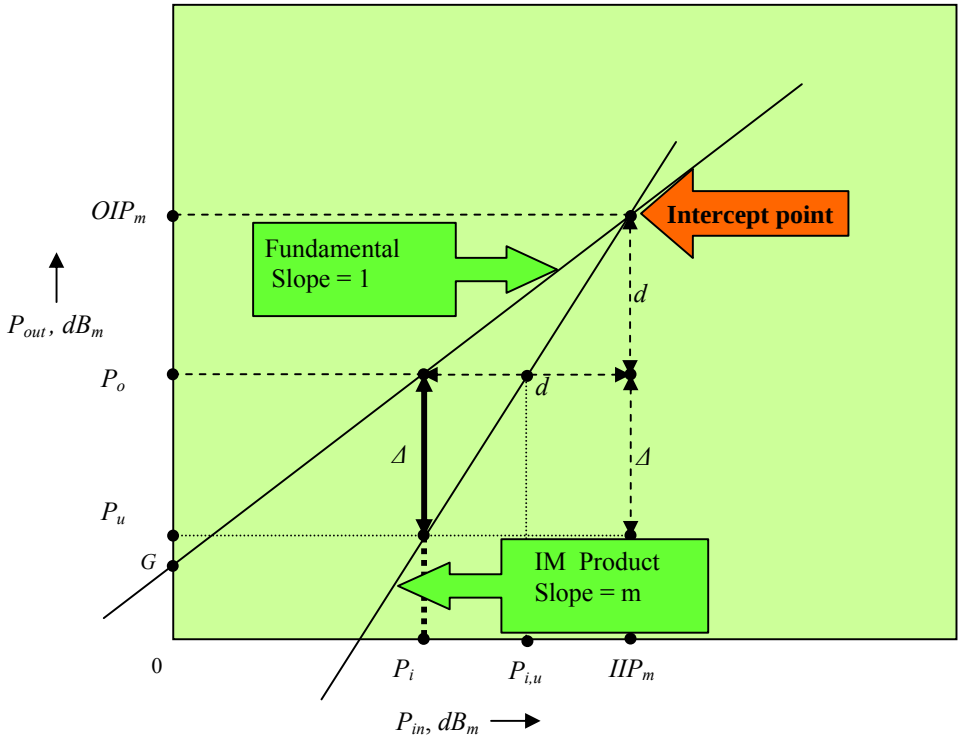
and, in the mixer design, the 2nd order inter-modulation is due to the spurious product with

$$m = 2 \quad , \quad \text{and} \quad n = 2 \quad , \tag{10.8}$$

which are called half IF spurs or Able-Baker spurs.

10.2 IP (Intercept point) and IMR (Inter-Modulation Rejection)

Intercept point of the 1st order signal power and the m^{th} order *IM* (Inter-Modulation) product power is an intermediate parameter to measure the linearity of a block or a system, by which the *IMR* of a block or a system can be calculated.



P_i = Input power of signal in *dB*,
P_o = Output power of signal in *dB*,
P_u = Undesired output power due to m^{th} non-linearity in *dB*,
G = Power gain of the device, or block, or system in *dB*,
OIP_m = The *m*th order output intercept point in *dB*,
IIP_m = The *m*th order input intercept point in *dB*,
d = Power difference between OIP_m and P_o in *dB*,
Δ = Output power difference between the 1st order and the m^{th} order in *dB*.

Figure 10.5 Plot of 1st order signal and m^{th} order *IM* product on the input-output power plane.

Figure 10.5 is a plot of 1st order signal and m^{th} order IM product on the input-output power plane. The straight line with slope of 1 depicts the relation between input and output signal power. The output power is G dB_m if the input power is 0 dB_m, where G is the power gain of the device or system. Another line with slope of m reflects the m^{th} order

IM (Inter-Modulation) product or spurious product. When the input signal power P_i is low, the output power of the m^{th} order IM product P_u is still much lower than the output power of the 1st order output signal P_o. Let Δ denote the output power difference between the 1st order and the m^{th} order when input signal power is P_i. The increase of the m^{th} order IM product output is much faster than that of the 1st order output signal power as the input signal power is increased from P_i, and eventually the two lines intercept as shown in Figure 10.5. This point is called the intercept point of the 1st order signal and the m^{th} order IM product. The corresponding coordinates are called the m^{th} order output intercept point OIP_m and the m^{th} order input intercept point IIP_m respectively.

For m^{th} order intercept point, it is easy to find out the relationship between the input intercept point IIP_m and the input signal power P_i. From the simple geometry derivation,

$$\Delta = m(IIP_m - P_i) - 1(IIP_m - P_i) = (m-1)(IIP_m - P_i) \ , \qquad (10.9)$$

$$IIP_m = \frac{\Delta}{(m-1)} + P_i \ , \qquad (10.10)$$

$$OIP_m = G + IIP_m \ . \qquad (10.11)$$

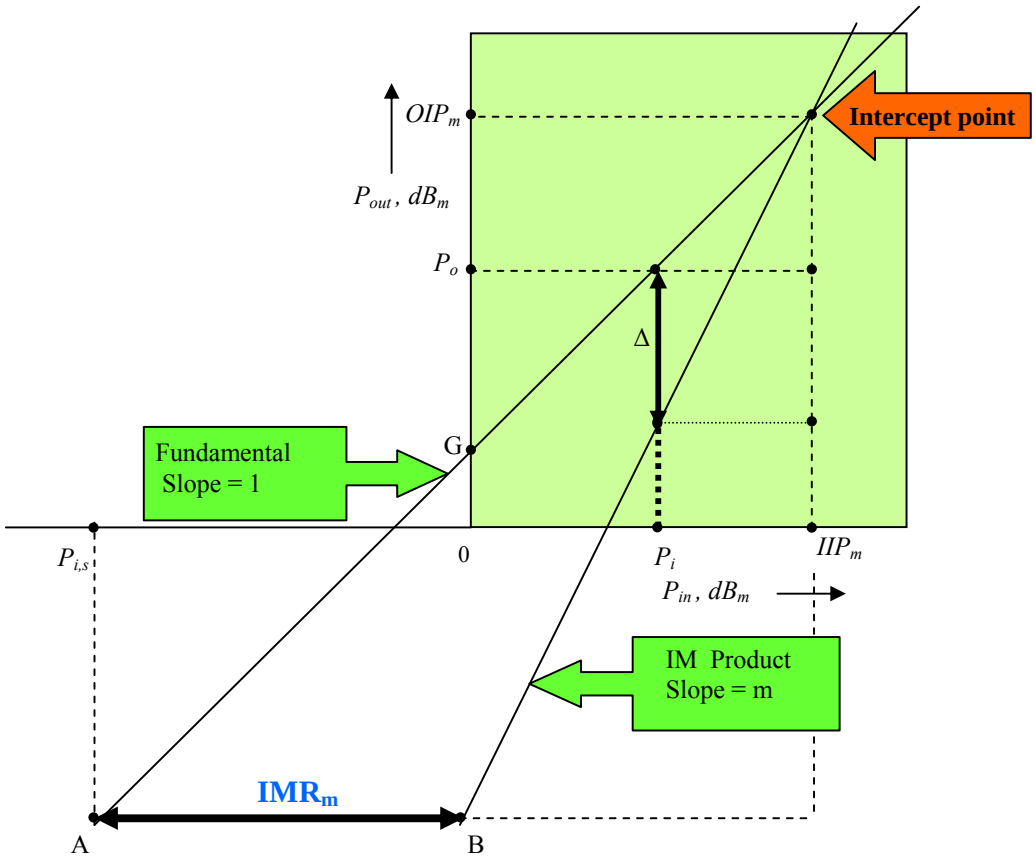


Figure 10.6 Calculation of IMR_m from intecept point, IIP_m.

By means of equation (10.10), the intercept point, either IIP_m or OIP_m, can be obtained by the measurement of Δ at an appropriate value of P_i.

Now, let's calculate the m^{th} order inter-modulation rejection, IMR_m, from the m^{th} order input intercept point, IIP_m. Figure 10.6 illustrates the meaning of the mth order inter-modulation rejection, IMR_m, which is expressed by the length of AB. In Figure 10.6, $P_{i,s}$ denotes the input power at the sensitivity point, the minimum of a detectable input signal level. Point A represents that the input signal power is at the sensitivity point and its corresponding output signal power. Point B is on the m^{th} order IM product line with the same output power as point A. The input power at point B is IMR_m dB_m higher than point A. In other words, when the input signal is at the sensitivity point, the m^{th} order IM product in the input is IMR_m dB_m stronger than the input signal power. As both input powers are increased towards IIP_m, they have the same output at the intercept point. Therefore, it implies that this system has a capability to suppress or to reject the m^{th} order IM product down (IMR_m) dB.

From the simple geometry as shown in Figure 10.6, we have

$$IMR_m = \left(IIP_m - P_{i,s}\right) - \frac{\left(IIP_m - P_{i,s}\right)}{m} \quad ,$$

$$IMR_m = \alpha\left(IIP_m - P_{i,s}\right) , \tag{10.12}$$

where

$$\alpha = \frac{(m-1)}{m} \ . \tag{10.13}$$

10.3 3rd Order Intercept Point and Spurious Product

As mentioned above, the power or amplitude of an IM or spurious product is relatively reduced as its order is increased. Therefore, more attention is paid on the IM or spurious products with lower orders. The IM or spurious products with m=2 or 3 have being more focused than others since the IM or spurious products with m=1 is the desired signal. In this section let's discuss the 3rd order spurious product.

When

$$m = 3, \tag{10.15}$$

then, expressions (10.10) to (10.13) become

$$IIP_3 = \frac{\Delta}{2} + P_i, \tag{10.16}$$

$$OIP_3 = G + IIP_3, \tag{10.17}$$

$$\alpha = \frac{2}{3}, \tag{10.18}$$

$$IMR_3 = \alpha\left(IIP_3 - P_{i,s}\right) = \frac{2}{3}\left(IIP_3 - P_{i,s}\right). \tag{10.19}$$

By means of a direct measurement of Δ in the test laboratory, one can calculate IIP_3 and then IMR_3 by equations (10.16) and (10.19).

On the other hand, in terms of the transfer function of the device or system, (10.4), one can evaluate the IMR_3. Assuming that the ω_{u1} and ω_{u2} in equation (10.5) are

$$\omega_{u1} = \omega_d - \delta\omega, \quad \text{or} \quad f_{u1} = f_d - \delta f, \tag{10.20}$$

$$\omega_{u2} = \omega_d - 2\delta\omega, \quad \text{or} \quad f_{u2} = f_d - 2\delta f, \tag{10.21}$$

where ω_{u1} or f_{u1} = undesired angular frequency or frequency,
ω_{u2} or f_{u2} = undesired angular frequency or frequency,
ω_d or f_d = desired angular frequency or frequency.

Note that

$$\omega = \omega_d = 2\omega_{u1} - \omega_{u2}, \tag{10.22}$$

Let's emphasize the term with the frequency, ω_d=$2\omega_1$-ω_2 in the transfer function of a non-linear device or system, the equation (10.4) can be re-written as

$$v_{out} = ... + [(3/4)\, a_3 v_o^{\,3} + ...]cos(2\omega_1 - \omega_2)t + ... \qquad (10.23)$$

This is the IM product by the two un-desired signals with frequencies ω_{u1} or ω_{u2} and with the same amplitude V_o. The measurement of IM rejection is to raise these two un-desired signals' levels up to a reference level as the same output is produced on channel:

$$v_{out} = ... + [a_1 v_s + ...]cos\omega_d t + ... \qquad (10.24)$$

where v_s = voltage of the desired signal. The term with $cos(2\omega_{u1} - \omega_{u2})t$ in expression (10.23) corresponding to an on-channel desired signal with $cos\omega_d t$ in expression (10.24), that is,

$$a_1 v_s \cos\omega_d t \approx \frac{3}{4} a_3 v_o^{\,3} \cos(2\omega_{u1} - \omega_{u2})t \quad , \qquad (10.25)$$

then,

$$v_o \approx \left(\frac{4}{3}\frac{a_1}{a_3} v_s\right)^{1/3} \quad , \qquad (10.26)$$

Consequently,

$$IMR_{3,v} = \frac{v_o}{v_s} \approx \left(\frac{4}{3}\frac{a_1}{a_3}\right)^{1/3} v_s^{\,-2/3} \quad , \qquad 10.27)$$

$$IMR_{3,dB} = 20\log\frac{v_o}{v_s} \approx 20\log\left[\left(\frac{4}{3}\frac{a_1}{a_3}\right)^{1/3} v_s^{\,-2/3}\right] . \qquad (10.28)$$

The 3rd order inter-modulation rejection, IMR_3, is not only dependent of the ratio of non-linearity coefficients, a_1 and a_3, but also sensitive to the input signal voltage, v_s. The higher signal voltage, the less the IMR_3.

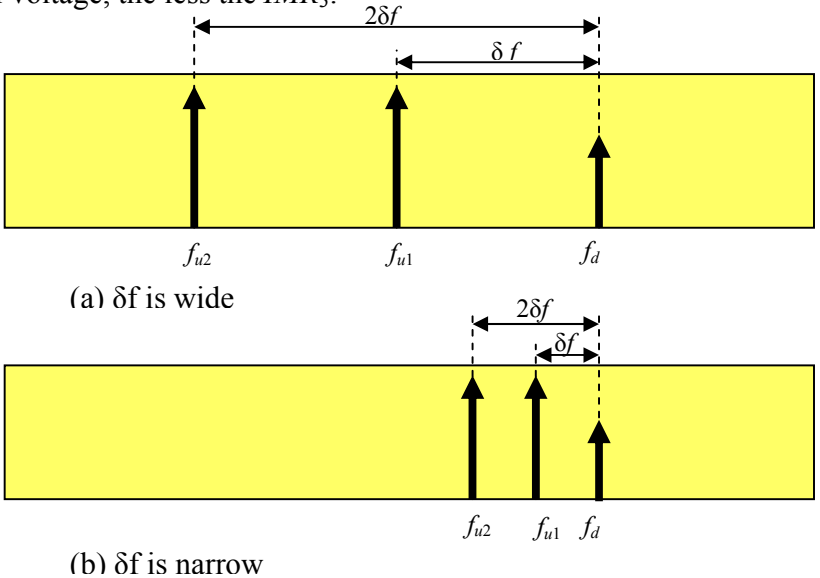


Figure 10.7 Undesired signal frequencies, f_{u1} and f_{u2}, and desired signal frequency, f_d.

Among many orders of the non-linearity, the 3rd order, IP_3 or IRM_3, is adapted as the main criterion of the non-linearity of a device or a system. Also, in the RF/RFIC circuit design, the engineer must take care of the 3rd order, IP_3 or IRM_3, more than other orders.

Why?

Let's return to the expressions of the two un-desired input signal frequencies, (10.20) and (10.21), and the desired signal frequency, (10.22). They are plotted in Figure 10.7.

An outstanding feature of the 3rd order spurious product is that the difference of the two un-desired frequencies, δf, could be any value! The δf could be several hundred *MHz* or 1 *Hz*. It implies that such kinds of IM or spurious products can never be removed away by means of a filter to be built. To build a narrow band filter with its bandwidth of 1 *Hz* or a few *Hz* is impossible practically while a wide band filter with several 100 *MHz* is usually meaningless in a wireless communication system. It implies that the 3rd order spurious can never be completely eliminated by means of any filtering technology thought it can be removed away in a limited bandwidth range. This feature of the 3rd order spurious is unique because other orders of spurious could be completely filtered out practically.

Theoretically, the non-linearity of a device or a system could be described by any order of IM or spurious product. However, by means of the unique feature of the 3rd order spurious product, the 3rd order intercept point, IP_3, is chosen as a criterion to judge the non-linearity of a device or a system if an appropriate low value of δf is taken in the measurement of a device or a system. For example, there are two cellular phones produced by Motorola and Nokia respectively. Both radio phones are operating in the same operating frequency range and have the same channel spacing or bandwidth, say,

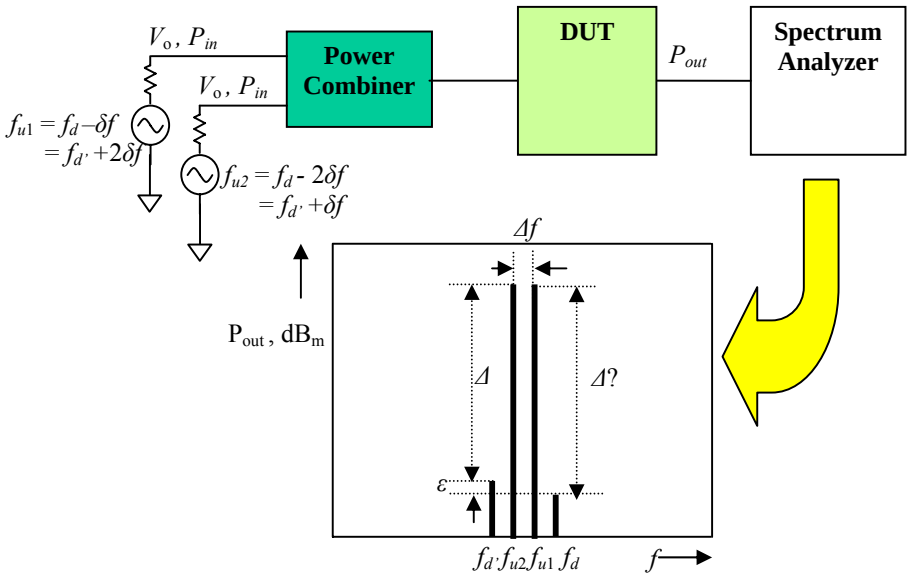


Figure 10.8 Set-up and display of IP_3

12.5 *kHz*. To compare their non-linearity performance without prejudice, FCC could choose any value of δf, which is less than 6.25 *KHz*, so that both of the two un-desired signals are totally located within the channel bandwidth and never be filtered away. The test result would be in justice to both of Motorola and Nokia without prejudice.

The set-up and display of IP_3 testing are shown in Figure 10.8. There are two generators generating the same power level, V_o or P_{in}, but with two frequencies, f_{u1} and f_{u2}. Two input signals are combined together by a Power Combiner and then delivered into the desired tested block or system. The power output, P_{out}, from the desired tested block or system is displayed on the screen of the spectrum analyzer. Not only product at f_d but also product at $f_{d'}$ can be seen because there is another 3rd order IM or spurious product symmetrically produced by the same sources, f_{u1} and f_{u2}. As a matter of fact, this product at $f_{d'}$ is corresponding terms with the frequency of $(2\omega_{u2}- \omega_{u1})$ in the expression (10.4).

Similar to (10.20), (10.21), and (10.22), the relationships between $f_{d'}$ and f_{u1}, f_{u2} are

$$\omega_{u1} = \omega_{d'} + 2\delta\omega \quad , \qquad \text{or} \qquad f_{u1} = f_{d'} + 2\delta f \quad , \tag{10.29}$$

$$\omega_{u2} = \omega_{d'} + \delta\omega \quad , \qquad \text{or} \qquad f_{u2} = f_{d'} + \delta f \quad , \tag{10.30}$$

$$\omega = \omega_{d'} = 2\omega_{u2} - \omega_{u1} \, , \tag{10.31}$$

In terms of equations (10.16) and (10.19), the values of IP_3 or IMR_3 can be calculated from the measurement of Δ.

There is a minor problem that happens quite often in the actual testing. Two values of Δ at f_d and $f_{d'}$ appeared on the display screen are not the same but with a difference, ε. This is due to the large number of computation iterations in the computer system and results from the truncation of the decimal number. The worst case resides with the lesser Δ, and should be taken into the calculation of IP or IMR.

10.4 1 dB Compression Point and IP_3

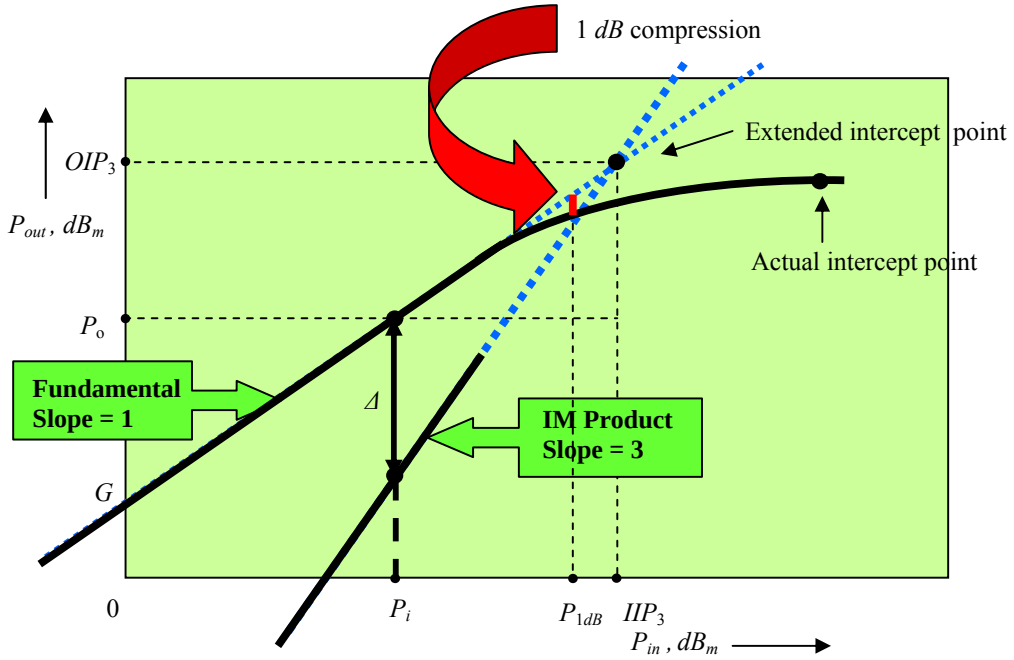


Figure 10.9 1 *dB* compression point and *IP*3

In the previous section the 3rd order intercept point for input and output, IIP_3 and OIP_3, is plotted in Figure 10.5. It is, in fact, a linear approximation of an actual measurement only. The actual testing result can be represented by the Figure 10.9, in which the plots are two straight lines with slopes of m=1 or m=3 respectively when the input power is low but start to bend downwards when the input power is increased up to a certain level. The plots then deviate from a straight line more and more as the input power continuously increases. The two straight lines can be extended in the increasing direction of input power. The extended intercept point of these two straight lines with slopes of m=1 and m=3 is the 3rd order intercept point, IIP_3 or OIP_3, that we are looking for. The two bended curves have an actual intercept point with a higher IIP_3 but lower OIP_3.

At the curve which contains the line with a lope of m=1, the bending point is not too far from the intercept point. The bending point is correlated with the 3rd order intercept point. The testing for the 3rd intercept point is quite complicated than the testing only for the curve containing the line with a slope of m=1. It would be simple to stuck on the testing only the curve containing the line with a slope of m=1 if the 3rd order intercept point can be evaluated from the bending point. Instead of the bending point, the 1 *dB* compression point is utilized to evaluate the 3rd order intercept point. As shown in Figure 10.9, the 1 *dB* compression point is a special input power, P_{1dB}, at which the output power deviates 1 *dB* from the extended straight line with a slope of m=1. The relationship between 1 *dB*, P_{1dB}, and 3rd order input intercept point, IIP_3, can be evaluated as

$$IIP_3 \rightarrow P_1 dB + (3 \text{ to } 10)\ dB, \tag{10.32}$$

The uncertain factor, (3 to 10) *dB*, depends on the type of device, circuit topology, current drain, DC voltage supply and other parameters of the device or system.

By the way, in the 3rd order intercept point testing, in order to cover enough linear portion, the input power should be

$$P_i < -30\ dB_m, \tag{10.33}$$

so as to ensure that the plot can cover enough linear portion and can be extended from the linear portion to get the 3rd order intercept point.

10.5 2nd order Intercept Point and Spurious Product

When

$$m = 2 \quad , \tag{10.34}$$

expressions (10.10) to (10.13) become

$$IIP_2 = \Delta + P_i \quad , \tag{10.35}$$

$$OIP_2 = G + IIP_2 \quad , \tag{10.36}$$

$$\alpha = \frac{1}{2} \, , \tag{10.37}$$

$$IMR_2 = \alpha\left(IIP_2 - P_{i,s}\right) = \frac{1}{2}\left(IIP_2 - P_{i,s}\right) . \tag{10.38}$$

By means of the direct measurement of Δ in the test laboratory, one can calculate IIP_2 and then IMR_2 by equation (10.35) and (10.38).

On the other hand, in terms of the transfer function of the device or system, (10.2), one can evaluate the IMR_2. Assuming that the ω in equation (10.2) is

$$\omega_u = \frac{\omega_d}{2} , \qquad \text{or} \qquad f_u = \frac{f_d}{2} , \tag{10.39}$$

where ω_u or f_u = undesired angular frequency or frequency,
ω_d or f_d = desired angular frequency or frequency.

The transfer function or v_{out} (10.2) contains the term with the frequency, $\omega_d = 2\omega_u$, is

$$v_{out} = \ldots + [(1/2)\, a_2 v_o^{\,2} + \ldots] cos 2\omega_u t + \ldots \tag{10.40}$$

This is the IM product by the un-desired signal with frequencies ω_u and with the same amplitude v_o. The measurement of IM rejection is to raise this un-desired signals' level up to a reference level as the same output that is produced on channel:

$$v_{out} = \ldots + [a_1 v_s + \ldots] cos\omega_d t + \ldots \tag{10.41}$$

where v_s = voltage of the desired signal. The term with $cos 2\omega_u t$ in expression (10.40) corresponds to an on-channel desired signal with $cos\omega_d t$ in expression (10.1), that is,

$$a_1 v_s \cos\omega_d t \approx \frac{1}{2} a_2 v_o^{\,2} \cos 2\omega_u t \, , \tag{10.42}$$

then,

$$v_o \approx \left(2\frac{a_1}{a_2} v_s \right)^{1/2}, \tag{10.43}$$

Consequently,

$$IMR_{2,v} = \frac{v_o}{v_s} \approx \left(2\frac{a_1}{a_2} \right)^{1/2} v_s^{-1/2}, \tag{10.44}$$

$$IMR_{2,dB} = 20\log\frac{v_o}{v_s} \approx 20\log\left[\left(2\frac{a_1}{a_2} \right)^{1/2} v_s^{-1/2} \right]. \tag{10.45}$$

The 2nd order inter-modulation rejection, IMR_2, is dependent not only on the ratio of non-linearity coefficients, a_1 and a_2, but also sensitive to the input signal voltage, v_s. The higher signal voltage, the less the IMR_2.

The set-up and display of IP_2 testing is shown in Figure 10.10. There are two generators generating the same power level, v_o or P_{in}, but with two frequencies f_d and f_u. Two input signals are toggle-switched by a switch. The power output P_{out} from the desired tested block or system is displayed on the screen of the spectrum analyzer.

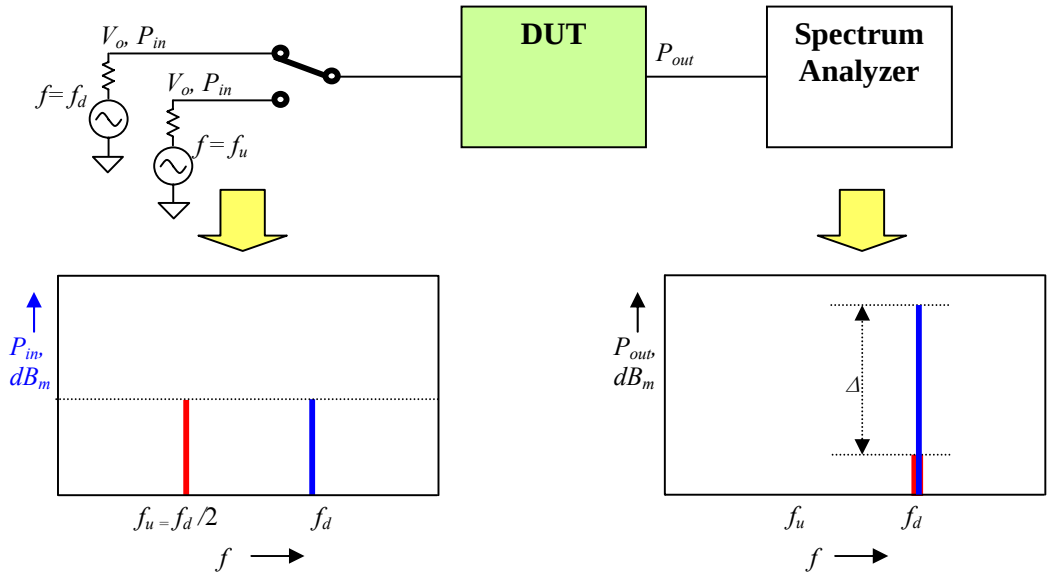


Figure 10.10 Set-up and display of IP_2 testing

In terms of equations (10.35) and (10.38), the values of IP_2 or IMR_2 can be calculated from the measurement of Δ.

10.6 Distortion

A distortion is an intuitive parameter and measured by percentage. For instance, a distorted RF voltage signal with a sinusoidal waveform implies that it is not a perfect signal but has an un-perfect portion, in which either the amplitude or the phase deviates from its ideal state. The deviation from the ideal state is measured by percentage, or *dB*. They are defined as

$$D_{v,\%} = \frac{\Delta V}{V} \quad , \tag{10.46}$$

$$D_{P,dB} = 20\log\frac{\Delta V}{V} \approx 20\log\left(D_{v,\%}\right) . \tag{10.47}$$

where $D_v,\%$ = Distortion of voltage in *%*,
$D_{P,dB}$ = Distortion in *dB*.

In terms of relation (10.47), distortion by perentage and by *dB* can be exchanged. For example, 5% of distortion is equal to about –26 *dB*:

$$(-26.02)dB = 20\log\left(\frac{5}{100}\right) . \tag{10.48}$$

It should be noted that the mechanism of the distortion and the noise for an RF signal is essentially different. The noise spreads or swallows its trace of waveform but does not bring about the distortion, while all of the harmonics and all IM or spurious products contribute to distortion but not to noise.

References

[1] F. F. Fulton, “Two-Tone Nonlinearity Testing - The Intercept Point P/sub i/,” Microwave Symposium Digest, G-MTT International , Volume: 73 , Issue: 1 , pp. 112 – 112, June 1973.

[2] Richard C. Sagers, “Intercept Point and Undesired Responses,” 32nd IEEE Vehicular Technology Conference, May 23-25, 1982.

[3] B. P. Lathi, “The Design of CMOS Radio-Frequency Integrated Circuits,” (Book), Cambridge University Press, 1983.

[4] Jack Smith, “Modern Communication Circuits,” (Book), McGraw-Hill Publishing Company, 1986

[5] George D. Vendelin, Anthony M. Pavio, and Ulrich L. Rohde, “Microwave Circuit Design Using Linear and Nonlinear Techniques,” (Book), John Wiley & Sons, Inc., 1990.

[6] Paul R. Gray, and Robert G. Meyer, “Analysis and Design of Analog Integrated Circuits,” (Book), Third Edition, John Wiley & Sons, Inc., 1993.

[7] T. S. Chu, “Intermodulation in CDMA,” Personal, Indoor and Mobile Radio Communications, 1994. Wireless Networks - Catching the Mobile Future. 5th IEEE International Symposium on , Vol. 2, pp. 595 – 600, 18-23 September 1994 .

[8] Keng Leong Fong, R. G. Meyer, “High-frequency nonlinearity analysis of common-emitter and differential-pair transconductance stages,” Solid-State Circuits, IEEE Journal of , Vol. 33 , Issue: 4 , pp. 548 – 555, April 1998.

[9] B. Bastani, E. E. Bautista, G. Nagaraj, J. Heck, “A quadrature down converter for direct conversion receivers with high 2nd and 3rd order intercept points,” Devices, Circuits and Systems, 2000. Proceedings of the 2000 Third IEEE International Caracas Conference on , pp. C19/1 - C19/4, 15-17 March 2000.

[10] E. E. Bautista, B. Bastani, J. Heck, “A high IIP2 downconversion mixer using dynamic matching,” Solid-State Circuits, IEEE Journal of , Vol. 35 , Issue: 12 , pp. 1934 – 1941, December 2000.

[11] K. Kivekas, A. Parssinen, K. A. I. Halonen, “Characterization of IIP2 and DC-offsets in transconductance mixers,” Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on [see also Circuits and Systems II: Express Briefs, IEEE Transactions on] , Vol. 48 , Issue: 11 , pp. 1028 – 1038, November 2001.

[12] K. Kivekas, A. Parssinen, and K. A. I. Halonen, “Characterization of IIP2 and DC-offsets in transconductance mixers,” Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on [see also Circuits and Systems II: Express Briefs, IEEE Transactions on] , Vol. 48 , Issue: 11 , pp. 1028 – 1038, Nov. 2001.

[13] C. R. Iversen, and T. E. Kolding, “Noise and intercept point calculation for modern radio receiver planning ,” Communications, IEE Proceedings- , Vol. 148 , Issue: 4 , pp. 255 – 259, August 2001.

[14] A, Ignea, R. Kortvelyessy, and O. Stanescu, “A new approach on nonlinearity distortion,” Telecommunications in Modern Satellite, Cable and Broadcasting Service, 2001. TELSIKS 2001. 5th International Conference on , Vol. 1 , pp. 301 - 304 , 19-21 September 2001.

[15] H. –D. Wohlmuth, W. Simburger, “ A high-IP3 RF receiver chip set for mobile radio base stations up to 2 GHz,” Solid-State Circuits, IEEE Journal of , Volume: 36 , Issue: 7 , pp. 1132 – 1137, July

2001.

[16] Min Lin, Yongming Li, and Hongyi Chen, "A novel IP3 boosting technique using feedforward distortion cancellation method for 5 GHz CMOS LNA," Microwave Symposium Digest, 2003 IEEE MTT-S International , Vol. 1 , pp. A185 - A188 , 8-13 June 2003.

[17] A. R. Petrov, "System approach for low 1/f noise, high IP2 dynamic range CMOS mixer design," University/Government/Industry Microelectronics Symposium, 2003. Proceedings of the 15th Biennial , pp. 74 – 77, 30 June-2 July 2003.

[18] A. A. Abidi, "General relations between IP2, IP3, and offsets in differential circuits and the effects of feedback," Microwave Theory and Techniques, IEEE Transactions on , Vol. 51 , Issue: 5 , pp. 1610 – 1612, May 2003.

[19] D. Manstretta, M. Brandolini, and F. Svelto, "Second-order intermodulation mechanisms in CMOS downconverters," Solid-State Circuits, IEEE Journal of , Volume: 38 , Issue: 3 , pp. 394 – 406, March 2003.

[20] Sanghoon Kang, and Bumman Kim, "Second order nonlinearity analysis of Gilbert mixer," Microwave Symposium Digest, 2003 IEEE MTT-S International , Vol. 1 , pp. A45 - A48, 8-13 June 2003.

[21] Sanghoon Kang, and Bumman Kim, "Second order nonlinearity analysis of Gilbert mixer," Radio Frequency Integrated Circuits (RFIC) Symposium, 2003 IEEE , pp. 559 – 562, 8-10 June 2003.

Index

Contents

Chapter 11 Cascaded Equations and System Analysis

Usually an RF system consists of a number of blocks. The system parameters must be calculated from the corresponding parameters of all the individual blocks. The basic skill to do this is how to calculate the system parameters from the corresponding parameters of two individual blocks as shown in Figure 11.1. As long as we master such a skill for two blocks, it can be extended to a system constructed with more than two blocks.

11.1 Cascaded Equation for Power Gain

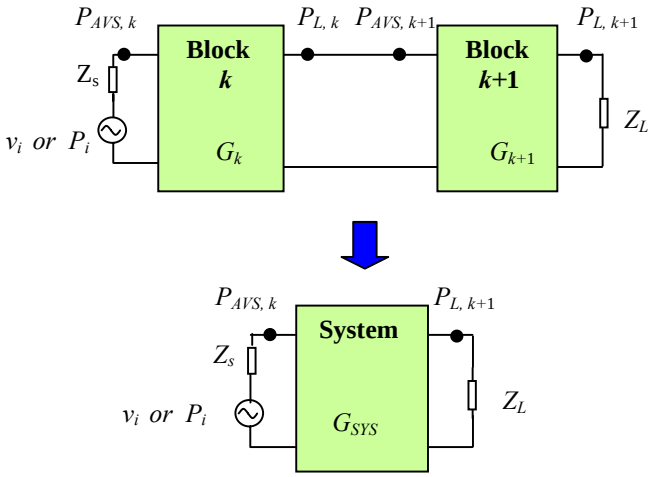


Figure 11.1 The system gain cascaded by two gains of block k and k+1.

Let's discuss a system containing of only two blocks as shown in Figure 11.1, where G_k and G_{k+1} are the transducer power gains of block k and k+1 respectively. From the discussion in Chapter 9, they are

$$G_k = G_{T,k} = \frac{P_{L,k}}{P_{AVS,k}} \quad , \tag{11.1}$$

$$G_{k+1} = G_{T,k+1} = \frac{P_{L,k+1}}{P_{AVS,k+1}} \quad , \tag{11.2}$$

then the system transducer power gain, G_{SYS}, should be

$$G_{SYS} = G_{T,SYS} = \frac{P_{L,k+1}}{P_{AVS,k}} = \frac{P_{L,k}}{P_{AVS,k}} \frac{P_{L,k+1}}{P_{L,k}} \quad , \tag{11.3}$$

where $P_{L,k}$ = Power delivered to the load of kth block,
$P_{L,k+1}$= Power delivered to the load of (k+1)th block,
$P_{AVS,k}$= Power available from the source of kth block,
$P_{AVS,k+1}$= Power available from the source of (k+1)th block,
G_k = Power gain of kth block,
$G_{T,k}$ = Transducer power gain of kth block,
G_{k+1}= Power gain of (k+1)th block,
$G_{T,k+1}$= Transducer power gain of (k+1)th block,
G_{SYS}= Power gain of system,
$G_{T,SYS}$= Transducer power gain of system.

Assuming that the impedances between block k and block k+1 are well-matched, it means that

$$P_{L,k} = P_{AVS,k+1} \quad . \tag{11.4}$$

Substituting (11.4) into (11.3), we have

$$G_{SYS} = G_{T,SYS} = \frac{P_{L,k}}{P_{AVS,k}} \frac{P_{L,k+1}}{P_{AVS,k+1}} \quad , \tag{11.5}$$

In terms of equations (11.1) and (11.2), in the numeric form or by the ratio of watts,

$$G_{SYS,watt} = G_k G_{k+1} \quad , \tag{11.6}$$

and, in the logarithic form or by the unit of dB, it becomes

$$G_{SYS,dB} = G_k + G_{k+1} \quad . \tag{11.7}$$

The system gain, G_{SYS}, is a simple product of the individual gains, G_k and G_{k+1}, in the numeric form, or $G_{SYS,dB}$, is a simple sum of the individual gains, G_k and G_{k+1}, in the logarithic form.

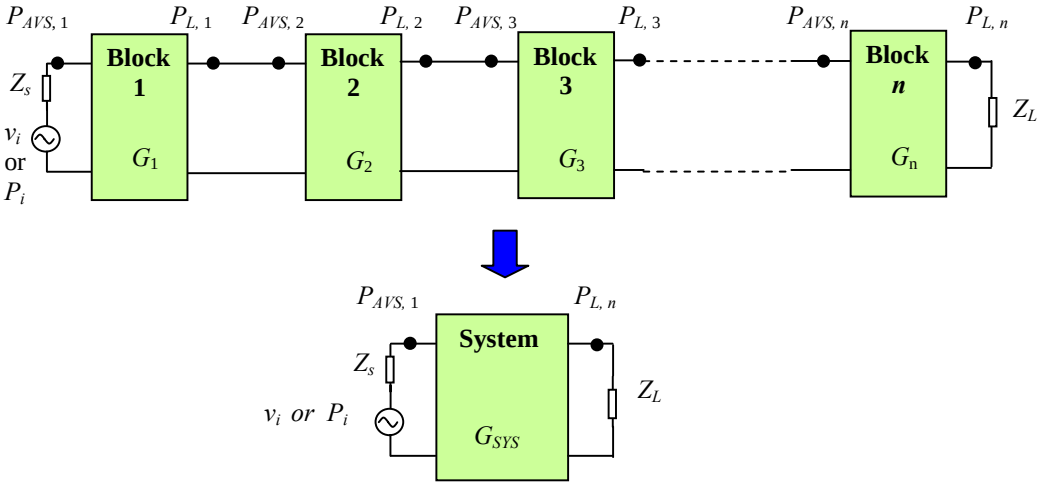


Figure 11.2 The system gain cascaded by individual gains from block 1 to block n.

In the general cases when the number of blocks is *n*, as shown in Figure 11.2, equations (11.6) and (11.7) can be extended to

$$G_{SYS} = G_1 G_2 G_3 .... G_n \quad , \tag{11.8}$$

$$G_{SYS,dB} = G_1 + G_2 + G_3 + ... + G_n \quad , \tag{11.9}$$

or,

$$G_{SYS} = \prod_1^n G_k \quad , \tag{11.10}$$

$$G_{SYS,dB} = \sum_1^n G_k \quad . \tag{11.11}$$

The expression (11.10) or (11.11) is the cascaded equation of the power gain.

11.2 Cascaded Equation for Noise Figure

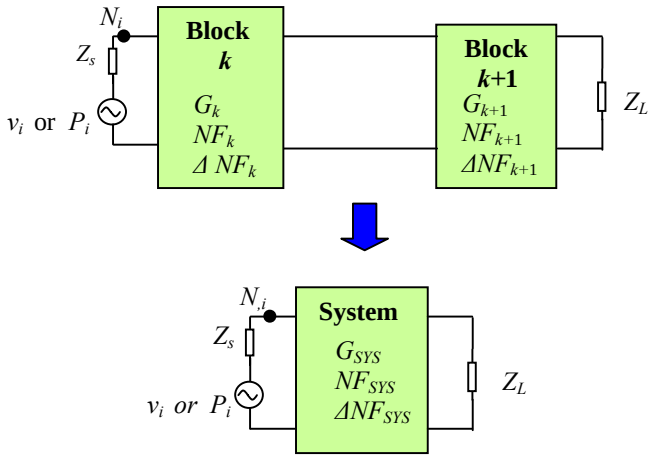


Figure 11.3 The system noise figure cascaded by two noise figures of block k and k+1.

Like the previous section, let's cascade two blocks into a resultant block or system. Figure 11.3 plots two blocks to be cascaded and the resultant block or system. Up to the definition of noise figure, the resulting noise figure is a ratio of the total output power to the output power due to the input noise power, that is,

$$NF_k = \frac{G_k N_i + \Delta N_k}{G_k N_i} = 1 + \frac{\Delta N_k}{G_k N_i} \ , \tag{11.12}$$

$$NF_{k+1} = \frac{G_{k+1} N_i + \Delta N_{k+1}}{G_{k+1} N_i} = 1 + \frac{\Delta N_{k+1}}{G_{k+1} N_i} \ . \tag{11.13}$$

Then,

$$NF_{SYS} = \frac{G_{k+1}\left(G_k N_i + \Delta N_k\right) + \Delta N_{k+1}}{G_k G_{k+1} N_i} = 1 + \frac{G_{k+1}\Delta N_k + \Delta N_{k+1}}{G_k G_{k+1} N_i} \ . \tag{11.14}$$

where $N_{,i}$ = Noise source power delivered to kth block,
ΔN_k = Additional noise power of kth block,
ΔN_{k+1}= Additional noise power of (k+1)th block,
ΔN_{sys}= Additional noise power of system,
G_k = Power gain of kth block,
G_{k+1}= Power gain of (k+1)th block,
NF_k = Noise figure of kth block,
NF_{k+1} = Noise figure of (k+1)th block,
NF_{SYS} = Noise figure of system in the numeric form.

Substituting (11.12) and (11.13) into (11.14), we have

$$NF_{SYS} = 1 + \frac{G_{k+1}(NF_k - 1)G_k N_i + (NF_{k+1} - 1)G_{k+1} N_i}{G_k G_{k+1} N_i}, \qquad (11.15)$$

$$NF_{SYS} = NF_k + \frac{(NF_{k+1} - 1)}{G_k}. \qquad (11.16)$$

The system noise figure, NF_{SYS}, in the numeric form or by the unit of watts, is not only a composed function of the individual noise figures, NF_k and NF_{k+1}, but also a function of the individual gain G_k of first block. From equation (11.16) it can be found that there is an important feature of the noise figure. The system noise figure or resulting noise figure NF_{sys}, consists of two terms. The first term NF_k is contributed by the first block k. The second term $(NF_{k+1}-1)/G_k$, is contributed by the second block $k+1$ but it is reduced down by a factor G_k. It means that, from the system viewpoint, the noise figure of the first block plays a more important role than that of the second one, while the noise figure of the second block can be reduced by the gain of first block if the gain of the first block is positive and appreciable. Consequently, in order to minimize the noise figure of system or to enhance the sensitivity of a receiver, the first important object to design the first block with low noise figure but high gain. This is the main goal for an LNA design in a receiver because the LNA is always located in the most front end of the receiver.

Assuming that a small system consists of two blocks, the LNA as block 1 and the filter as block 2, by means of cascaded noise figure and gain equations described above, the same system noise figure can be reached with different combinations of individual noise figures and gains as shown in Figure 11.4 and Table 11.1. The LNA has the same noise figure but different gain in the two examples. The gain of the LNA in example 2 is 6 *dB* higher than that in example 1. The second block, a filter, has 4 *dB* more insertion loss in example 2 than in example 1. It implies that the alternative approaches to the same sensitivity exist. The extra insertion loss in block 2 can be compensated by increasing the gain in block 1. It therefore provides flexibility in the design for the receiver's front end.

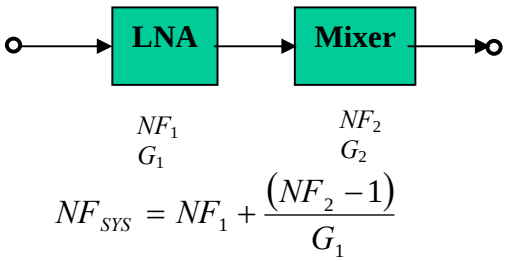


Figure 11.4 System noise figure if a system is cascaded by block 1 and 2.

Table 11.1 Identical system noise figure contributed from two blocks

with different settlements of noise figure and gain.

Example #1

	Block 1 LNA	Block 2 FLT	System
NF, dB	**2**	**3**	**2.17**
Gain, dB	**12**	**-3**	**9.00**

Example #2

	Block 1 LNA	Block 2 FLT	System
NF, dB	**2**	**7**	**2.17**
Gain, dB	**18**	**-7**	**11.00**

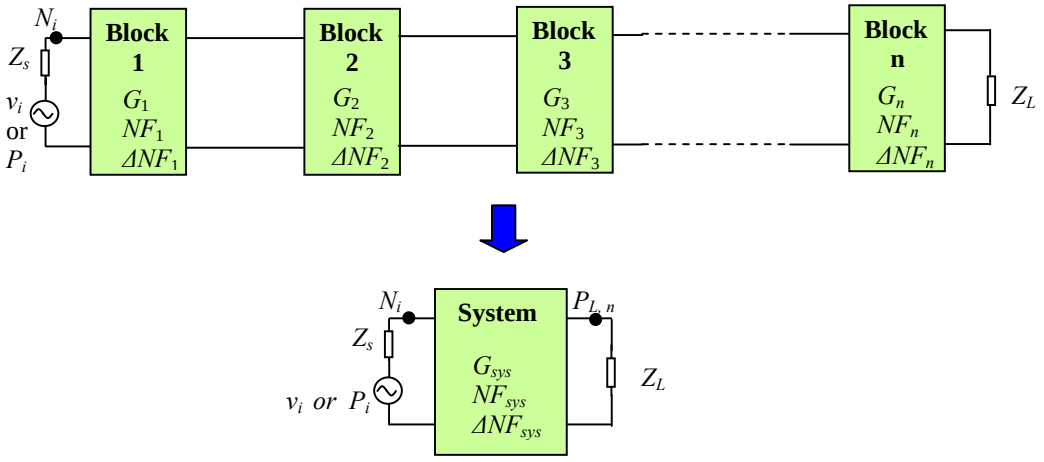


Figure 11.5 The system noise figure cascaded by individual noise figures from block 1 to block *n*.

In general cases when the number of blocks is *n*, as shown in Figure 11.5, equation (11.16) can be extended to

$$NF_{SYS,watt} = NF_1 + \frac{(NF_2 - 1)}{G_1} + \frac{(NF_3 - 1)}{G_1 G_2} + \frac{(NF_4 - 1)}{G_1 G_2 G_3} + ... + \frac{(NF_n - 1)}{G_1 G_2 \ldots G_{n-1}} \quad , \qquad (11.17)$$

or,

$$NF_{SYS,watt} = NF_1 + \sum_{2}^{n} \frac{(NF_k - 1)}{\prod_{1}^{k-1} G_j} \quad . \qquad (11.18)$$

The expression (11.17) or (11.18) is the cascaded equation of the noise figure.

11.3 Cascaded Equation for Intercept Point

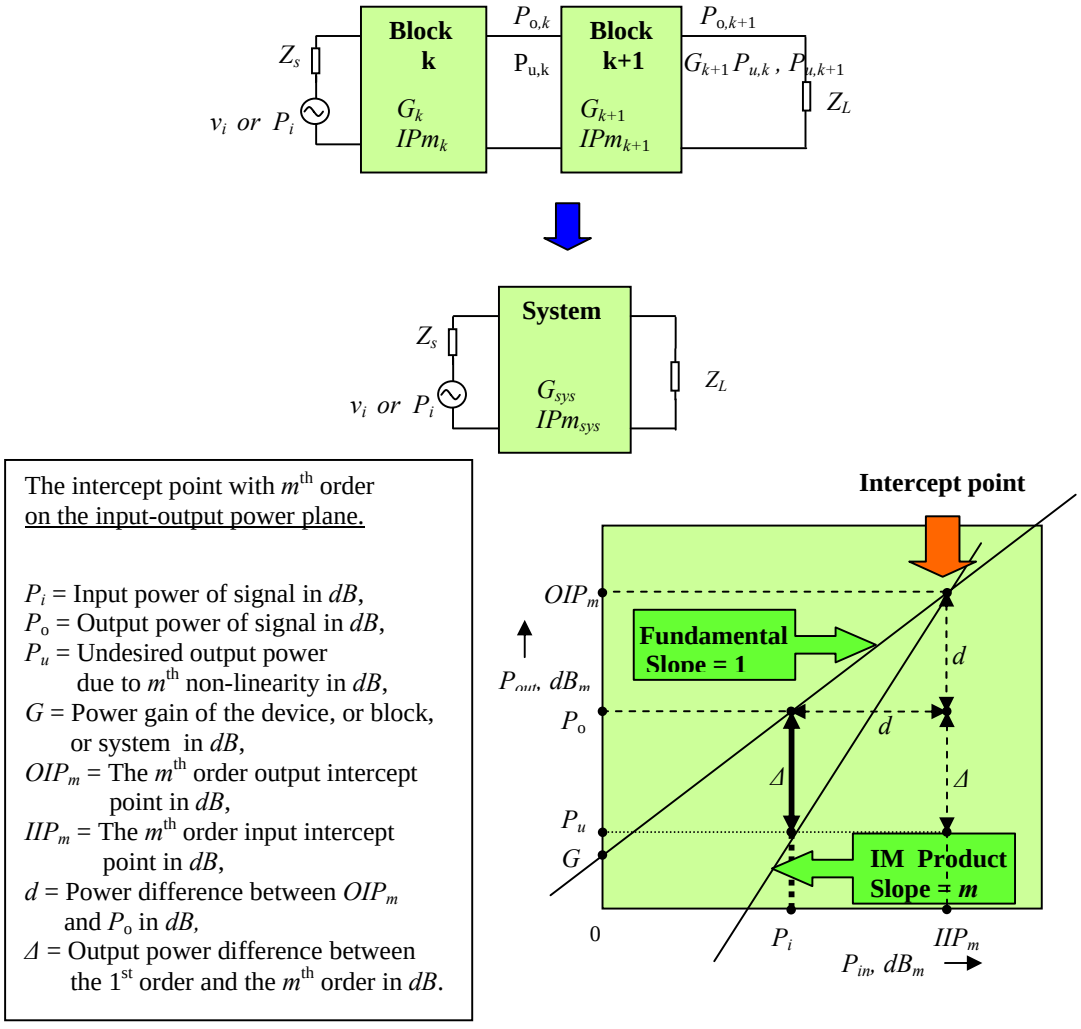


Figure 11.6 The system IP_m cascaded by two IP_ms of block n and n+1.

Figure 11.6 shows a system IP_m cascaded by two individual IP_ms.

From the simple geometry derivation, we have

$$\Delta = d(m-1) \quad . \tag{11.19}$$

Note that

$$\Delta = P_o - P_u \quad , \tag{11.20}$$

$$d = OIP_m - P_o \quad , \tag{11.21}$$

we have

$$P_u = P_o - \Delta = P_o - d(m-1) = P_o - (OIP_m - P_o)(m-1), \tag{11.22}$$

$$P_u = mP_o - (m-1)OIP_m, \tag{11.23}$$

In the numeric form or by the unit of *watt*s, equation (11.23) becomes

$$P_u = \frac{{P_o}^m}{OIP_m{}^{(m-1)}}, \tag{11.24}$$

Referring to the individual blocks, k and k+1, in Figure 2.21, equation (11.24) can be re-written as

$$P_{u,k} = \frac{{P_{o,k}}^m}{OIP_{m,k}{}^{(m-1)}}, \tag{11.25}$$

$$P_{u,k+1} = \frac{{P_{o,k+1}}^m}{OIP_{m,k+1}{}^{(m-1)}}, \tag{11.26}$$

where $P_{o,k}$ = Desired output power from k^{th} block,
$P_{o,k+1}$ = Desired output power from $(k+1)^{\text{th}}$ block,
$P_{u,k}$ = Undesired output power from k^{th} block,
$P_{u,k+1}$ = Undesired output power from $(k+1)^{\text{th}}$ block,
$OIP_{m,k}$ = Output intercept point from k^{th} block,
$OIP_{m,k+1}$ = Output intercept point from $(k+1)^{\text{th}}$ block,

Note that

$$P_{o,k} = \frac{P_{o,k+1}}{G_{k+1}}, \tag{11.27}$$

then

$${P_{o,k}}^m = \frac{{P_{o,k+1}}^m}{{G_{k+1}}^m}. \tag{11.28}$$

The undesired output power at Z_L from block k is

$$P_{u,k}G_{k+1} = \frac{{P_{o,k}}^m G_{k+1}}{OIP_{m,k}{}^{(m-1)}} = \frac{{P_{o,k+1}}^m G_{k+1}}{{G_{k+1}}^m OIP_{m,k}{}^{(m-1)}} = \frac{{P_{o,k+1}}^m}{{G_{k+1}}^{(m-1)} OIP_{m,k}{}^{(m-1)}}. \tag{11.29}$$

The undesired voltage at Z_L from block k is

$$V_{u,k} = \left[\frac{{P_{o,k+1}}^m Z_L}{{G_{k+1}}^{(m-1)} OIP_{m,k}{}^{(m-1)}} \right]^{\frac{1}{2}}. \tag{11.30}$$

From equation (11.26), the undesired voltage at Z_L from block k+1 is

$$V_{u,k+1} = \left[\frac{P_{o,k+1}{}^{m} Z_L}{OIP_{m,k+1}^{(m-1)}} \right]^{\frac{1}{2}}. \tag{11.31}$$

Then, the total undesired voltage or the system undesired voltage across Z_L is

$$V_{u,sys} = V_{u,k} + V_{u,k+1} = \sqrt{P_{o,k+1}{}^{m} Z_L} \left[\left(\frac{1}{G_{k+1} OIP_{m,k}} \right)^{\frac{(m-1)}{2}} + \left(\frac{1}{OIP_{m,k+1}} \right)^{\frac{(m-1)}{2}} \right]. \tag{11.32}$$

The total undesired output power or the system undesired output power across Z_L is

$$P_{u,sys} = \frac{V_{u,sys}{}^{2}}{Z_L} = P_{o,k+1}{}^{m} \left[\left(\frac{1}{G_{k+1} OIP_{m,k}} \right)^{\frac{(m-1)}{2}} + \left(\frac{1}{OIP_{m,k+1}} \right)^{\frac{(m-1)}{2}} \right]^2. \tag{11.33}$$

$$\frac{P_{u,sys}}{P_{o,k+1}} = P_{o,k+1}{}^{(m-1)} \left[\left(\frac{1}{G_{k+1} OIP_{m,k}} \right)^{\frac{(m-1)}{2}} + \left(\frac{1}{OIP_{m,k+1}} \right)^{\frac{(m-1)}{2}} \right]^2. \tag{11.34}$$

At the system intercept point,

$$P_{u,sys} = P_{o,k+1}, \tag{11.35}$$

$$P_{o,k+1} = OIP_{m,T}, \tag{11.36}$$

then, equation (11.34) becomes

$$\left(\frac{1}{OIP_{m,sys}} \right)^{\frac{(m-1)}{2}} = \left(\frac{1}{G_{k+1} OIP_{m,k}} \right)^{\frac{(m-1)}{2}} + \left(\frac{1}{OIP_{m,k+1}} \right)^{\frac{(m-1)}{2}}. \tag{11.37}$$

Mathematically, equation (11.37) looks like a very nice form of the relationship between the total output intercept point, $OIP_{m,sys}$, and the individual output intercept point, $OIP_{m,k}$ and $OIP_{m,k+1}$. The total output intercept point, $OIP_{m,sys}$, is not only a function of the individual output intercept point, $OIP_{m,k}$ and $OIP_{m,k+1}$, but also is related to the gain of the second individual block, G_{k+1}.

Now let's transfer equation (11.37) from output intercept point, OIP_m, to input intercept point, IIP_m. Note that

$$IIP_{m,sys} = \frac{OIP_{m,sys}}{G_T} = \frac{OIP_{m,sys}}{G_k G_{k+1}}, \tag{11.38}$$

also,

$$IIP_{m,k} = \frac{OIP_{m,k}}{G_k} \quad , \tag{11.39}$$

and,

$$IIP_{m,k+1} = \frac{OIP_{m,k+1}}{G_{k+1}} \quad , \tag{11.40}$$

then, we have

$$\left(\frac{1}{IIP_{m,sys}}\right)^{\frac{(m-1)}{2}} = \left(\frac{1}{IIP_{m,k}}\right)^{\frac{(m-1)}{2}} + \left(\frac{G_k}{IIP_{m,k+1}}\right)^{\frac{(m-1)}{2}}. \tag{11.41}$$

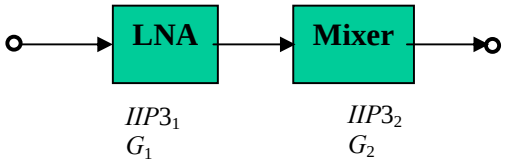
It can be re-written as

$$\left(\frac{1}{IIP_{m,sys}}\right)^{\frac{(m-1)}{2}} = \left(\frac{1}{IIP_{m,k}}\right)^{\frac{(m-1)}{2}} + \left(\frac{1}{\frac{IIP_{m,k+1}}{G_k}}\right)^{\frac{(m-1)}{2}}.$$

The system intercept point $IIP_{SYS,watt}$ in the numeric form or in the unit of watts, is not only a composed function of the individual intercept point $IIP_{m,k}$ and $IIP_{m,k+1}$, but also a function of the individual gain G_k of first block. The system intercept point or resulting intercept point $IIP_{m,sys}$ consists of two terms. The first term $(IIP_{m,k})^{-(m-1)/2}$ is contributed by the first block k . The second term, $(IIP_{m,k+1}/ G_k)^{-(m-1)/2}$, is contributed by the second block k+1 but it is reduced down by a factor G_k. The second term has more impact to the system intercept point if the gain of first block, G_k, is positive and appreciable. Equation (11.41) looks like the equation which describes a resultant resistor from two resistors in parallel if we treat IIP as the resistance. The difference is that the value of the second "resistor" must be G_k times decreased.

In order to clarify this feature of intercept point, let's examine some examples of the calculated results of IIP_3 by equation (11.41), which are listed in Table 11.2. The LNA is taken as the first block denoted by k while the mixer is taken as the second block denoted by k+1. It is found that in all of examples the value of system IIP_3 is always lower than that of the individual block's IIP_3 due to its feature of "resistors in parallel". In the first example, the resulting IP_3 is contributed by the IP_3 of the LNA and mixer as if a resulting "resistor", IP_{sys}=-3.5 *dBm*, is combined by a "resistor" of the LNA, IP_3=2 *dBm*, and a reduced "resistor" of the mixer, $IP_3'=IP_3/G_1$=(10-12)*dBm*=-2 *dBm*, in parallel, where G_1 is the power gain of the first "resistor" of the LNA. In the combination of two resistors in

parallel, the resistor with smaller value plays more important role to the resulting resistor. This is why the value of resulting IP_3=-3.5 *dBm* is closer to the value of reduced IP_3'=-2 *dBm*. This feature is furthermore verified in the second example, in which the value of LNA's IP_3 is intentionally increased up to 100 *dBm*, the resulting IP_3=-2 *dBm* is almost entirely cotributed by the reduced IP_3 of mixer, $IP_3'=IP_3/G_1$=(10-12)*dBm*=-2 *dBm* and the contribution from the LNA IP_3=100 *dBm* is negligible. In the third example, the resulting IP_3=2 *dBm* is almost entirely cotributed by the LNA's $IP_3=2$ *dBm* and the contribution from the mixer, $IP_3'=IP_3/G_1$=(100-12)*dBm*=-88 *dBm*, is negligible. Finally, in the fourth example where $G_1=0$ *dB* and the "two resistors in parallel" have the same value of IP_3=2 *dBm*, they have the identical contribution to the resulting "resistor" and the resulting IP_3 is –1 *dBm*, which is 3 dB less than the individual value of 2 *dBm*.



$$\left(\frac{1}{IIP\ 3_{sys}}\right)=\left(\frac{1}{IIP\ 3_1}\right)+\left(\frac{G_1}{IIP\ 3_2}\right)$$

Figure 11.7 System *IIP*3 if a system is cascaded by block 1 and 2.

Table 11.2 System intercept point contributed from two blocks

	Block 1: LNA	Block 2 : Mixer	System
Example #1			
IIP3, dBm	**2**	**10**	**-3.5**
Gain, dB	**12**	**0**	**12.0**
Example #2			
IIP3, dBm	**100**	**10**	**-2.0**
Gain, dB	**12**	**0**	**12.0**
Example #3			
IIP3, dBm	**2**	**100**	**2.0**
Gain, dB	**12**	**0**	**12.0**
Example #4			
IIP3, dBm	**2**	**2**	**-1.0**
Gain, dB	**0**	**0**	**0**

The feature of cascaded intercept point is somehow upside down from that of cascaded noise figure. Unlike the noise figure, the intercept point of the first block does not dominates over the other block's intercept point. On the contrary, as we see in the two blocks' case above, in the contribution to the resulting intercept point, the second block with a smaller reduced intercept point dominates over the first block.

Now let's extend our discussion to a general case: The system consists of *n* blocks as shown in Figure 2.22. Equations (11.37) and (11.41), can be extended as

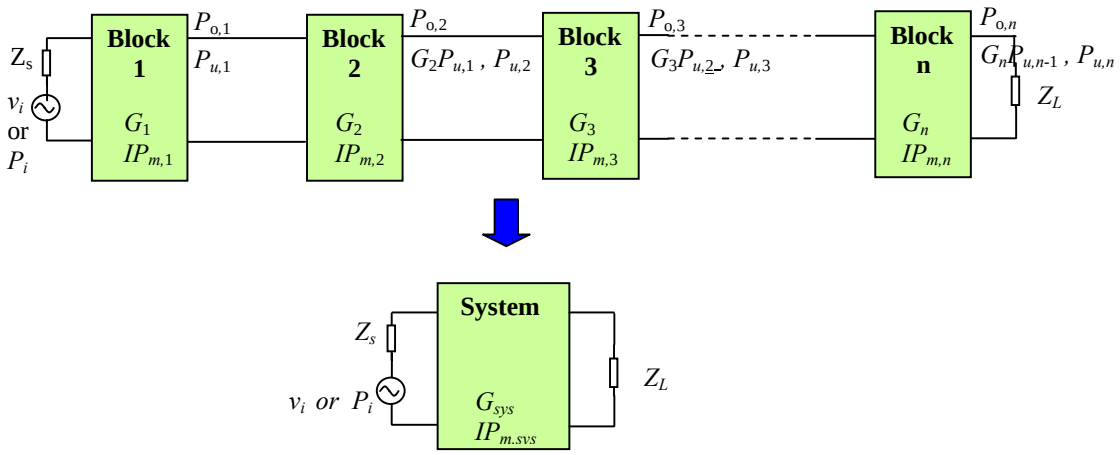


Figure 11.8 The system intercept point cascaded by individual intercept points from block 1 to block *n*.

$$\left(\frac{1}{OIP_{m,T}}\right)^{\frac{(m-1)}{2}}=\left(\frac{1}{OIP_{m,1}G_2G_3G_4..G_n}\right)^{\frac{(m-1)}{2}}+\left(\frac{1}{OIP_{m,2}G_3G_4..G_n}\right)^{\frac{(m-1)}{2}}+\cdots+\left(\frac{1}{OIP_{m,n-1}G_n}\right)^{\frac{(m-1)}{2}}+\left(\frac{1}{OIP_{m,n}}\right)^{\frac{(m-1)}{2}}. \tag{11.42}$$

$$\left(\frac{1}{IIP_{m,T}}\right)^{\frac{(m-1)}{2}}=\left(\frac{1}{IIP_{m,1}}\right)^{\frac{(m-1)}{2}}+\left(\frac{G_1}{IIP_{m,2}}\right)^{\frac{(m-1)}{2}}+\left(\frac{G_1G_2}{IIP_{m,3}}\right)^{\frac{(m-1)}{2}}+\cdots+\left(\frac{G_1G_2G_2\cdots G_{n-1}}{IIP_{m,n}}\right)^{\frac{(m-1)}{2}}. \tag{11.43}$$

Or,

$$\left(\frac{1}{OIP_{m,T}}\right)^{\frac{(m-1)}{2}}=\sum_{1}^{n-1}\left(\frac{1}{OIP_{m,k}\prod_{k+1}^{n}G_j}\right)^{\frac{(m-1)}{2}}+\left(\frac{1}{OIP_{m,n}}\right)^{\frac{(m-1)}{2}}, \tag{11.44}$$

$$\left(\frac{1}{IIP_{m,T}}\right)^{\frac{(m-1)}{2}} = \left(\frac{1}{IIP_{m,1}}\right)^{\frac{(m-1)}{2}} + \sum_{2}^{n}\left(\frac{\prod_{1}^{k-1} G_j}{IIP_{m,k}}\right)^{\frac{(m-1)}{2}} \quad . \qquad (11.45)$$

In the mathematic derivations above, all of the individual blocks are assumed to have flat frequency responses. It implies that the frequency response of a block to the input desired signal with frequency f_d and the input spurious with frequency f_s has the same gain, while the input spurious with frequency f_s, produces the undesired IM product with the frequency f_u. This is, of course, not true in a block with frequency selectivity, where the gain to the input spurious with frequency f_s, is not the same as the gain to the input desired signal with frequency f_d. The block k-1 as shown in Figure 11.9 is a block with an additional selectivity *SEL dB*, at the input spurious frequency f_s. It indicates that the gain at the input spurious frequency f_s is *SEL* dB lower than the gain at the input desired signal frequency f_d when both of them go through the block k-1.

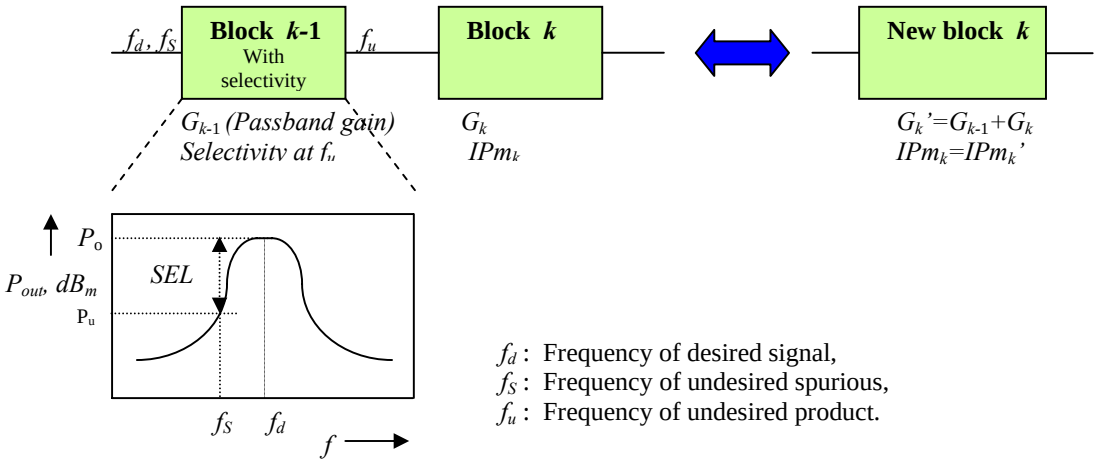


Figure 11.9 Combination of two blocks into one block when the selectivity is existed in the first one .

In such a case, block k-1 should be combined with the following block k to form a new block k. It is easy to understand that for the desired input signal,

$$G_k' = G_{k-1} + G_k \quad , \qquad (11.46)$$

where G_k' = The combined gain to the desired input signal in the new block k,
G_{k-1} = The gain to the desired input signal in the block k-1,
G_k = The gain to the desired input signal in the block k.

The impact of selectivity can be illustrated by Figure 11.10. Referring to geometry as shown in Figure 11.10, we have

$$P_{i,u} - P_i = d - \frac{d}{m} = \frac{m-1}{m} d = \frac{m-1}{m}\left(IIP_m - P_i\right) \quad . \tag{11.47}$$

The variation of $(P_{i,u}\text{-}P_i)$ in respect to IIP_m is

$$\delta\left(P_{i,u} - P_i\right) = \frac{m-1}{m}\delta\left(IIP_m\right) \quad , \tag{11.48}$$

The difference $(P_{i,u}\text{-}P_i)$ represents the input inter-modulation rejection IMR_i as shown in Figure 11.10 by the solid-arrow line. The selectivity of block k-1 enable $P_{i,u}$ pumped SEL dB up so that the outputs of the 1st and 3rd order are at the same level. It implies that the variation $\delta(P_{i,u}\text{-}P_i)$ or $\delta(IMR_i)$ is equal to SEL, that is,

$$\delta\left(P_{i,u} - P_i\right) = \delta\left(IMR_i\right) = \left(P_{i,u}{}' - P_{i,u}\right) = SEL \quad . \tag{11.49}$$

consequently, the straight line with a slope of m moves SEL dB toward the right side as shown in Figure 11.10 and the intercept point, IP_m, is moved to a new position, $IP_m{}'$.

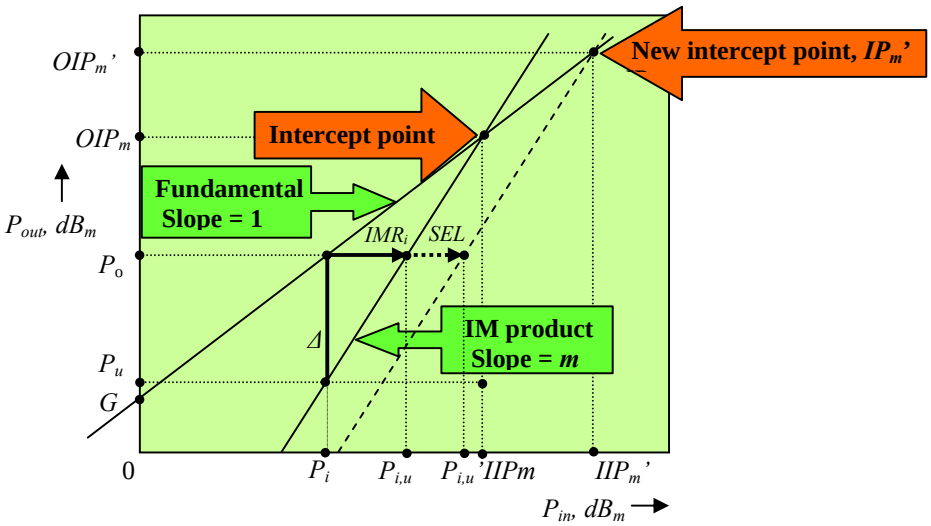


Figure 11.10 The relationship between IP_m and $IP_m{}'$ when $P_{i,u}$ is moved to $P_{i,u}{}'$.

From expressions (11.48) and (11.49), we have

$$\delta\left(IIP_m\right) = \frac{m}{m-1} SEL \quad . \tag{11.50}$$

Consequently, in the logarithmic form or by the unit of dB,

$$IIP_m{}' = IIP_m + \delta(IIP_m) = IIP_m + \frac{m}{m-1}SEL \quad , \qquad (11.51)$$

or, in the numeric form or in the unit of *watt*s,

$$IIP_m{}' = IIP_m SEL^{\frac{m}{m-1}} \quad . \qquad (11.52)$$

In summary, in the case of the block with selectivity, the block should be combined with the following block to form a new block as shown in Figure 11.11. The power gain of the new block to the desired input signal must be calculated from equation (11.46), and in the cascaded equations of input intercept point, IIP_m must be replaced by $IIP_m{}'$ as expressed in equation (11.52). Correspondingly, in the cascaded equations of output intercept point, OIP_m, must be replaced by $OIP_m{}'$, which can be derived from the relations, (11.39) and (11.52), that is,

$$OIP_m{}' = GIIP_m{}' = GIIP_m SEL^{\frac{m}{m-1}} \quad . \qquad (11.53)$$

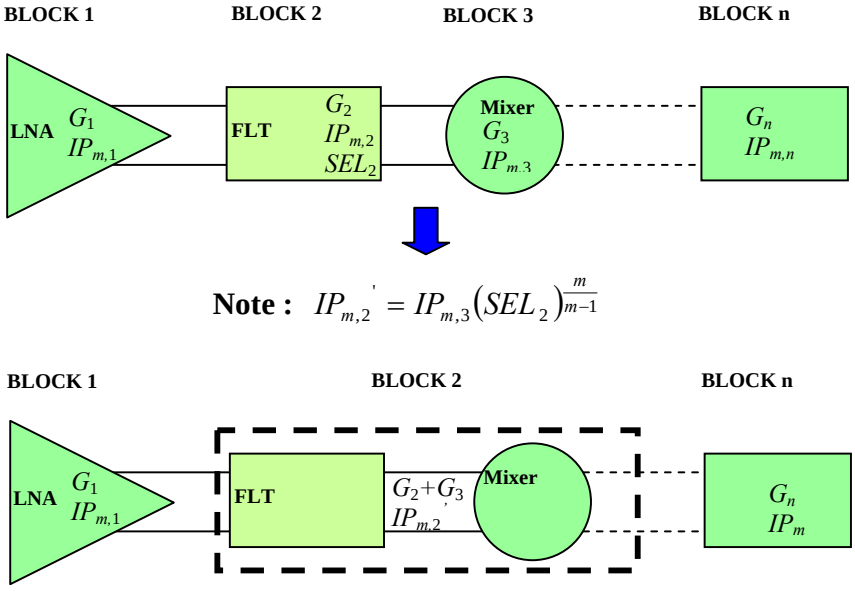


Figure 11.11 In the case of the block with selectivity, the block should be combined with the following block to form a new block.

11.4 Application of Cascaded Equations in the System Analysis

At present, the software for computer simulation offers a very convenient and sophisticated tool not only for the circuit design, but also for the system performance analysis. However, the cascaded equations enable engineers to do a system analysis in a fast and handy way, though such an analysis stays in a primary stage.

Table 11.3 System analysis of a receiver front end by the cascaded equations

System goals										
	Lower Frequency	: 403 *MHz*				Temperature	: 298 K^o			
	Upper Frequency	: 470 *MHz*				*RISE*@12*dBSINAD*	: 6.0 *dB*			
	IF Frequency	: 73.35 *MHz*				*RISE*@20*dB*Quiet	: 7.6 *dB*			
	System BW	: 13.5 *kHz*				Worst case Factor	: 0.1			
Performance										
(Column)		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
		Ant.SW	FLT#1	RF Amp.	FLT#2	Mixer	Xtal FLT1	IF Amp.	Xtal FLT2	Bk.End
***Gain*,**	***dB***	-0.9	-2	12	-1.5	-1	-2	10	-2	
***NF*,**	***dB***	0.9	2	2.5	1.5	7.5	2	3	2	6
***IP3*,**	***dB$_m$***	30	18	10	100	18	1000	12	1000	20
***IP2*,**	***dB$_m$***	100	100	100	100	48	1000	100	1000	1000
***SEL@Δf*,**	***dB***	0	0	0	0	0	17	0	17	0
***SEL@2Δf*,**	***dB***	0	0	0	0	0	29	0	29	0
***SEL@1/2* IF,**	***dB***	0	10	0	0	0	0	0	0	0
Calculations										
Gain (worst),	***dB***	-1	-2.2	10.8	-1.7	-1.1	-2.2	9.0	-2.2	
NF (worst),	***dB***	1.0	2.2	2.8	1.7	8.3	2.2	3.3	2.2	6.6
***NFsys*,**	***dB***	7.0	6.1	4.1	11.2	9.7	6.0	4.0	8.0	6.0
NFsys (worst)	***dB***	8.3	7.3	5.1	12.3	10.7	6.9	4.7	8.8	6.6
12dBSINAD	***uV***	0.20	0.18	0.15	0.33	0.28	0.18	0.14	0.23	0.18
12dBSINAD	***dB$_m$***	-120.8	-121.7	-123.7	-116.6	-118.1	-121.8	-123.8	-119.8	-121.8
***20dBQuiet*,**	***uV***	0.26	0.23	0.18	0.42	0.35	0.23	0.18	0.29	0.23
***20dBQuiet*,**	***dB$_m$***	-118.8	-119.7	-121.7	-114.6	-116.1	-119.8	-121.8	-117.8	-119.8
***IIP3sys*,**	***dB$_m$***	8.1	7.2	5.6	19.5	18.0	45.5	12.0	53.5	20.0
IMR3	***dB***	85.9	85.9	86.2	90.7	90.7	111.5	90.5	115.5	94.5
***IIP2sys*,**	***dB$_m$***	52.7	50.9	27.0	51.0	48.0	104.0	100.0	995.8	1000
IMR2	***dB***	86.8	86.3	75.3	83.8	83.0	112.9	111.9	557.8	560.9

Table 11.3 is an example to show the system analysis by the cascaded equations as discussed in this chapter. This is a front end of a receiver. It consists of 8 blocks. The 9th block in last column is the back end of receiver and is excluded and constructed in the base-band system.

The basic parameters of the system goals are listed in the upper portion of the Table. In the intermediate portion, the input data of performance for every individual block is presented. On the basis of the input data and by means of the cascaded equations of gain, noise figure, and intercept point, the calculations are conducted in the bottom portion.

The first two rows in the calculation portion are the values of gain and noise figure in the worst case, in which a 10% of degradation from the input values in the upper portion is adapted.

Instead of simultaneously for all the blocks, the calculations are confined for only two blocks. On other word, only cascaded equations for two blocks are applied in the calculations in Table 11.3.

Calculation starts from the last two blocks, the back end (Bk end) block and the crystal filter 2 (Xtal FLT2) block, or last two columns, 8 and 9. The input data in the back end block are provided from the Base-band system designer.

The cascaded block from the two blocks in columns 8 and 9 is temporarily called the "system" block. The 1st calculated "system" results are registered in the column 8, which would be treated as a 1st temporary block. The 2nd calculation is for the 1st temporary block cascaded with the IF amplifier block in column 7. The 2nd calculated "system" results are registered in column 7, which is treated as a 2nd temporary block. The 3rd calculation is for the 2nd temporary block cascaded with the IF amplifier in column 6. Again, the 3rd calculated "system" results are registered in column 6, which is treated as a 3rd temporary block, and so on. The calculation is iterated from right side to left side until the 1st block in column 1 is cascaded and the cascaded calculation is done.

Based on the calculated gain, noise figure, and intercept point, we can furthermore calculate the sensitivity, say, the 12 *dB SINAD*, the IMR (Inter-Modulation Rejection), and are also shown in the Table 11.3.

Table 11.3 shows a successful design since the final results would make the performance excellent :

12*dB,SINAD*	-120.8 dB_m
IMR_3	85.9 *dB*
IMR_2	86.8 *dB*

The calculations performed in Table 11.3 are quite flexible. If using Microsoft Excel spread sheet, the calculated results are automatically corrected if any input data is varied. It is very helpful to the system engineer in the adjustment of the design goals between the individual blocks.

References

[1] Richard C. Sagers, “Intercept Point and Undesired Responses,” 32nd IEEE Vehicular Technology Conference, May 23-25, 1982.

[2] Richard C. Sagers, “Intercept Point and Undesired Responses,” Transactions of IEEE Vehicular Technology, Vol. VT-32, No. 1, pp. 121-133, February, 1983.

[3] B. P. Lathi, “The Design of CMOS Radio-Frequency Integrated Circuits,” (Book), Cambridge University Press, 1983.

[4] Robert M. Gagliardi, “Satellite Communications,” (Book), Lifetime Learning Publications, 1984.

[5] G. Gonzales, “Microwave Transistor Amplifier Analysis and Design,” (Book), Prentice-Hall, 1984.

[6] Jack Smith, “Modern Communication Circuits,” (Book), McGraw-Hill Publishing Company, 1986

[7] Bernard Sklar, “Digital Communications,” (Book), Prentice Hall, 1988.

Index

Contents

Chapter 12 From Analog to Digital Communication System

The main difference between analog and digital communication systems is the difference of modulation and demodulation techniques. Figure 12.1 shows the basic components in both analog and digital modulation system respectively.

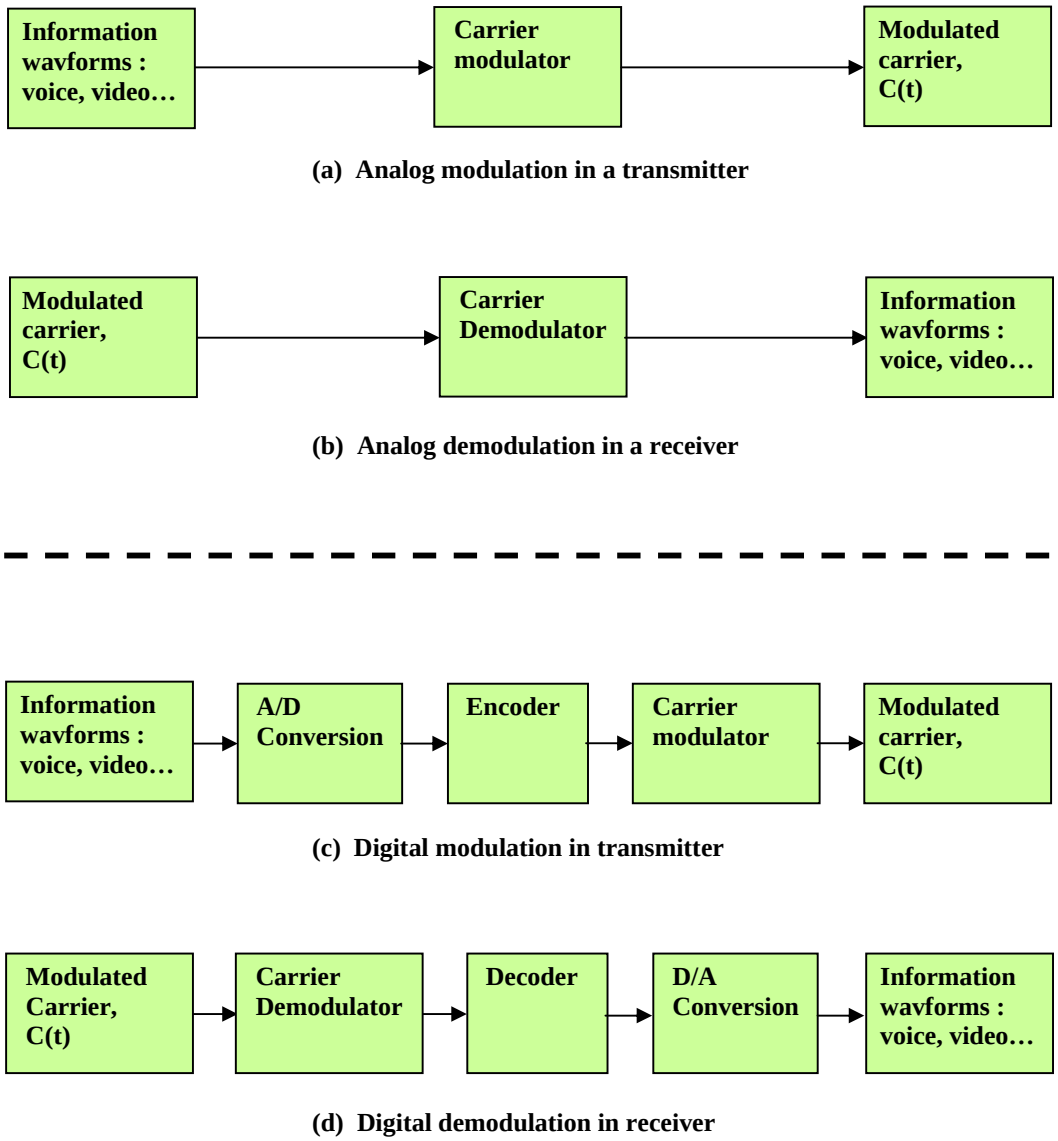


Figure 12.1 The basic components in both analog and digital modulation system

In a transmitter with analog modulation, the information waveforms, such as voice, video, and so on, are directly modulated from the source onto the carrier, while in a transmitter with digital modulation, the information waveforms are first converted into sequences of digital symbols by an A/D (Analog to Digital) processing, secondly encoded by a specific coding scheme, and finally modulated onto the carrier.

In a receiver with analog modulation, the information waveforms are directly demodulated from the carrier to the information waveforms, while in a receiver with digital modulation, the modulated carrier is first demodulated into the coding waveforms, secondly decoded into sequences of digital symbols, and finally converted into information waveforms by the D/A (Digital to Analog) processing.

The analog modulation techniques have been developed and predominated in early stage of communication systems. The general trend definitely favors digital communications. Compared with analog system, the main advantages of a digital communication are

1) Digital bits can be transmitted relatively error-free while using less carrier power than is required to operate an analog system;
2) Most hardware blocks in both the transmitter and receiver are digital circuits. In contrast to analog circuits, the digital circuits are simpler and thus easier to be implemented.

Compared with an analog system, the main disadvantage of a digital communication system is that A/D and D/A conversions must be added as shown in Figure 12.1. However, this disadvantage is diminishing by the continual development of high-speed, low-weight, and low-costing A/D and D/A conversion circuitry.

In this chapter we review the basic modulation technologies in both analog and digital communication system.

12.1 Modulation in an analog communication system

Basically, there are three types of modulation in analog communication systems: AM (Amplitude Modulation), FM (Frequency Modulation), and PM (Phase Modulation). As shown in Figure 12.1, in the transmitter the information waveforms are directly modulated onto the carrier by one of the AM, FM, or PM techniques while in the receiver the information waveforms are directly de-modulated from the modulated carrier by a corresponding demodulation technique. These three types of modulation can be described by their modulated carriers respectively:

For AM modulation,

$$C(t) = A[1 + \Delta_a m(t)]\cos(\omega_c t + \Psi) \ , \tag{12.1}$$

$$P_c = \frac{A^2}{2}(1 + \Delta_a{}^2 P_m) \ , \tag{12.2}$$

for FM modulation,

$$C(t) = A\cos(\omega_c t + 2\pi\Delta_f \int m(t)dt + \Psi) \ , \tag{12.3}$$

$$P_c = \frac{A^2}{2} \ , \tag{12.4}$$

and for PM modulation,

$$C(t) = A\cos(\omega_c t + \Delta_p m(t) + \Psi) \ , \tag{12.5}$$

$$P_c = \frac{A^2}{2} \ , \tag{12.6}$$

where C(t) = Modulated carrier,
A = Carrier amplitude,
Δ_a = AM modulation coefficient or AM index, *volt/volt*,
Δ_f = FM modulation coefficient or frequency deviation coefficient, *Hz/volt*,
Δ_p = PM modulation coefficient or phase deviation coefficient, rad/volt,
m(t) = Information waveform,
ω_c = Angular frequency of carrier,
Ψ = Phase angle of carrier,
P_c = power of carrier,
P_m = power in the information waveform.

The power of an AM modulated carrier is adjusted by the power in the information waveform. The modulated carrier with variable amplitude is therefore very easily disturbed by interferences from either the transceiver noise or noise in the propagation path. On the other hand, a modulated carrier with a non-constant envelope produces a profound non-linear distortion to the signal. In contrast to the AM modulation system, A FM or PM communication system has more capability to prevent from the interference by

the noises. Therefore, at present communication systems with AM modulation have been mostly replaced by communication systems with FM or PM modulation.

In a communication system, the key parameters of a modulated carrier are its bandwidth *BW*, the minimum of received *CNR* (Ratio of Carrier to Noise) before demodulation, and the *SNR* (Ratio of Signal to Noise) after demodulation. Table 12.1 lists these parameters for three types of modulation.

Table 12.1 Key parameters for three types of modulation

Modulation type	BW_n	Required *CNR*	Demodulated *SNR*
AM	$2BW_m$	> 12 *dB*	$=CNR$
FM	$2(\beta+1)BW_m$	> 16 *dB*	$6\beta^2(\beta+1)CNR$
PM	$2(\Delta_p+1)BW_m$	> 10 to 12 *dB*	$6\Delta_p^2(\Delta_p+1)CNR$

Notes: BW_n = Noise bandwidth ;
BW_m = Bandwidth of m(t);
β = Peak frequency deviation/BW_m;
Δ_p = Phase deviation.

The noise bandwidth BW_n is directly related to the bandwidth of m(t), BW_m. The bandwidth BW_m depends on the type of sources, on the number of sources and on the manner in which they are combined to form m(t). The bandwidth BW_m is the bandwidth of the source m(t) itself if only a single source is used as the modulating waveform in the channel which is the case of so-called SCPC (Single-Channel-Per-Carrier) carrier. Table 12.2 lists the frequency bandwidth of some sources BW_m.

Table 12.2 Bandwidth BW_m of some sources.

Type of source	BW_m
Voice	4 *kHz*
Multiplexed voice, 1000 voice sub-channels	4 *MHz*
Video	4 *MHz*
Television channel	6 *MHz*

The second row in Table 12.2 is simply equal to the summation of 1000 voice channels' bandwidth. Obviously, it corresponds to a form of FDM (Frequency Division Multiplexing).

The difference between *CNR* and *SNR* in Table 12.1 should be clarified. The carrier-to-noise ratio is usually to specify the status of the modulated carrier before de-modulation in the receiver or the status of the modulated carrier after the modulator in the transmitter. Especially, the demodulator of a receiver requires a minimum of *CNR* for a suitable demodulation. Such a minimum is called the threshold value of *CNR* for the receiver and is dependent of the specific type of the demodulator. The signal-to-noise ratio usually is to specify the status of the demodulated signal after de-modulation in the receiver or the status of the modulated carrier before the modulator in the transmitter. The value of *SNR*

depends on the specific modulation type as shown in Table 12.1 and on the modulator characteristics. For AM modulation, the *SNR* is almost equal to *CNR*. For FM and PM modulation, the *SNR* is usually higher than the *CNR* as shown in Table 12.1. Its enhanced factor is $6\beta^2(\beta+1)$ for FM modulation or $6\Delta_p^2(\Delta_p+1)$ for PM modulation. However, the price of this advantage is that the wider bandwidth must be provided. Compared with AM modulation, the enhanced factor of bandwidth is $(\beta+1)$ for FM modulation or (Δ_p+1) for PM modulation.

12.2 Encoding in a Digital Communication System

Encoding is a special element in digital communication systems. The information waveforms are first converted to sequences of data bits or binary symbols "0" and "1", by an A/D processing. The second step is encoding processing, that is, the data bits are specifically encoded to baseband waveforms by the encoder. The third step or final step in transmitter is to modulate the encoded baseband waveforms onto the carrier.

There are two kinds of encoding schemes: binary encoding and block encoding. In binary encoding, data bits are encoded bit by bit. In block encoding, data bits are encoded by groups of bits. Forward-error correction techniques for improving performance can be applied in these two encoding schemes by introducing redundant bits. Error correction can involve either fixed encoding or convolutional encoding. Of course, error correction increases the complexity of the entire encoding and decoding system.

The baseband waveform or the waveform after encoding can be expressed by

$$m(t) = \sum_{i=-\infty}^{\infty} d_i W(t - iT_b) \ , \tag{12.7}$$

where

$$d_i = \pm 1 \ , \tag{12.8}$$

where d_i = ith antipodal digit.
A digital data bit "1" is encoded into a specific waveform when d_i=1, while a digital data bit "0" is encoded into the negative of that waveform when d_i=-1.
W = a specific waveform representing a binary data bit "1",
t = time,
T_b = bit period.

The bit rate is the reciprocal of T_b, that is,

$$R_b = \frac{1}{T_b} \ , \text{ bits/sec.} \tag{12.9}$$

12.2.1 NRZ (Non-Return to Zero) and Manchester Format

There are two popular formats in the encoding processes: NRZ and Manchester. The NRZ format can be expressed by

$$\begin{aligned} W(t) &= 1 \ , \quad \text{when} \quad 0 \le t \le T_b \ , \\ &= 0 \ , \quad \text{elsewhere} \ . \end{aligned} \tag{12.10}$$

while the Manchester format can be expressed by

$$W(t) = 1, \qquad \text{when} \quad 0 \le t \le \frac{T_b}{2} \ ,$$

$$= -1, \qquad \text{when} \quad \frac{T_b}{2} \le t \le T_b \ . \tag{12.11}$$

The spectrum of the baseband signal, m(t), can be expressed by

$$S_m(\omega) = \frac{1}{T_b}\left|W(\omega)\right|^2 \ , \tag{12.12}$$

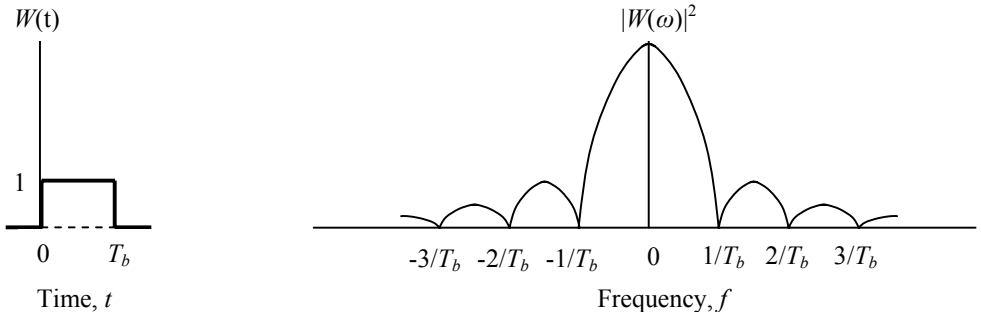
where $W(\omega)$ is the Fourier Transform of W(t).

The spectral density for NRZ format is

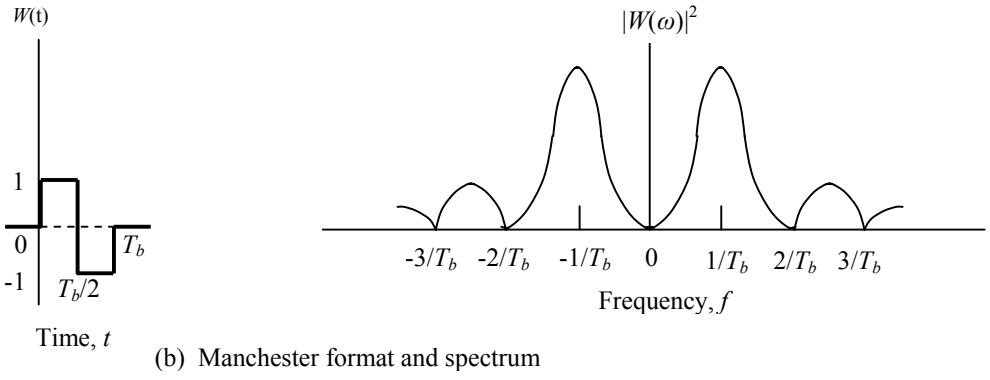
$$\left|W(\omega)\right|^2 = \left|\frac{\sin(\pi f T_b)}{\pi f T_b}\right|^2 \ , \tag{12.13}$$

and for Manchester format is

$$\left|W(\omega)\right|^2 = \left|\frac{\sin^4\left(\pi f \frac{T_b}{2}\right)}{\left(\pi f \frac{T_b}{2}\right)^2}\right|^2 \ , \tag{12.14}$$



(a) NRZ waveform and spectrum



(b) Manchester format and spectrum

Figure 12.2 Baseband signal and spectra

Figure 12.2 shows the baseband signals and the corresponding spectra.

The NRZ bit stream contains no sign change in a period. The advantage of an NRZ signal is that only positive DC power supply is required for its hardware circuitry. Another advantage is the narrow bandwidth of the main spectral humps, $2/T_b$ *Hz*, which is required to transmit the NRZ carrier. However, its main disadvantage is its low frequency interference caused by the waveform, which has a spectrum concentration around f=0.

The first advantage of Manchester waveform is its synchronization. Every Manchester bit contains a pulse sign change at exactly mid-bit time, which can be applied to establish a synchronized receiver clock at that precise bit period. The second advantage of Manchester waveform is the absence of low frequency interference because the Manchester waveform concentrates its spectrum around $f=1/T_b$, which is far away from f=0. The disadvantage is that both positive and negative DC power supplies are required. Another disadvantage is a wider bandwidth of the main spectral humps, $4/T_b$ *Hz*, which is required to transmit the Manchester carrier and is doubly wider than that required for the NRZ carrier.

12.2.2 BPSK (Binary Phase Shift Keying)

In BPSK (Binary Phase Shift Keying), the encoded baseband signal, m(t), is used to phase-modulate the carrier. The modulated phase $\Delta_p m(t)$ and expression (12.5) becomes

$$\Delta_p m(t) = \frac{\pi}{2}[1 - m(t)] = 0, \quad \text{when } m(t) = 1$$
$$= \pi, \quad \text{when } m(t) = -1. \qquad (12.15)$$

$$C(t) = A\cos(\omega_c t + \frac{\pi}{2}[1 - m(t)] + \Psi) \ . \qquad (12.16)$$

Then, in terms of the triangular formula $cos(a\text{-}b)=cos(a)cos(b)+sin(a)sin(b)$, because the value of the modulated phase is either 0 or π, we have

$$C(t) = Am(t)\cos(\omega_c t + \Psi) \ . \qquad (12.17)$$

This implies that the BPSK carrier is the same as an amplitude-modulated carrier in which the baseband signal, *m(t)*, directly multiplies the carrier. In other words, instead of phase modulation, the BPSK carrier can be formed directly by signal multiplication. It therefore simplifies the encoding circuitry significantly. Also, expression (12.17) implies that the resulting spectrum of the BPSK carrier is

$$S_c(\omega) = \frac{A^2}{4} S_m(\omega \pm \omega_c) \ . \qquad (12.18)$$

where $S_c(\omega)$= Spectrum of *BPSK* carrier;
$S_m(\omega)$= Spectrum of baseband waveform.

$S_m(\omega\pm\omega_c)$ represents the spectrum of baseband signal with a frequency shift of $\pm\omega_c$. Figure 12.3 shows $S_m(\omega+\omega_c)$, the spectrum of baseband signal with $+\omega_c$ frequency shifting.

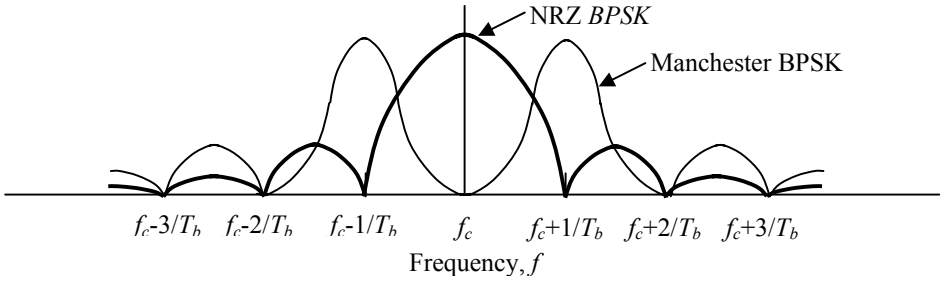


Figure 12.3 BPSK carrier spectra

With the exception of the main hump, there are many sub-humps or spectral tails in Figure 12.3. They might cause interference with other carriers though they are relatively lower than the main hump. It is therefore desirable to reduce the sub-humps or the spectral tails as much as possible. As a matter of fact, these sub-humps or the spectral tails are caused by the rectangular NRZ or Manchester waveform. It is thus interesting work to find a good waveform of which its spectral tails would decay much faster than produced by the rectangular waveform, while the constant carrier envelope is still maintained. One of the types of such a waveform is the RC (Raised Cosine) pulse as shown in Figure 12.4:

$$W(t)=\frac{1}{2}\left[1-\cos\left(\frac{2\pi t}{T_b}\right)\right], \qquad 0\le t\le T_b \tag{12.19}$$

The corresponding spectral density of an RC waveform (12.19) is

$$\left|W(\omega)\right|^2=\left|\frac{\sin(\pi fT_b)}{\pi fT_b}\right|^2\left|\frac{\cos(\pi fT_b)}{1-(2fT_b)^2}\right|^2 . \tag{12.20}$$

It can be found that when $fT_b<1$, the spectral density of the RC waveform, (12.20), approaches that of the NRZ waveform, (12.13), because the second factor in (12.20) approaches 1. This is why the two curves are close from each other in Figure 12.4 when $fT_b<1$. From Figure 12.4 it can be seen that the spectral tail of the SC waveform is significantly reduced from that of the NRZ waveform.

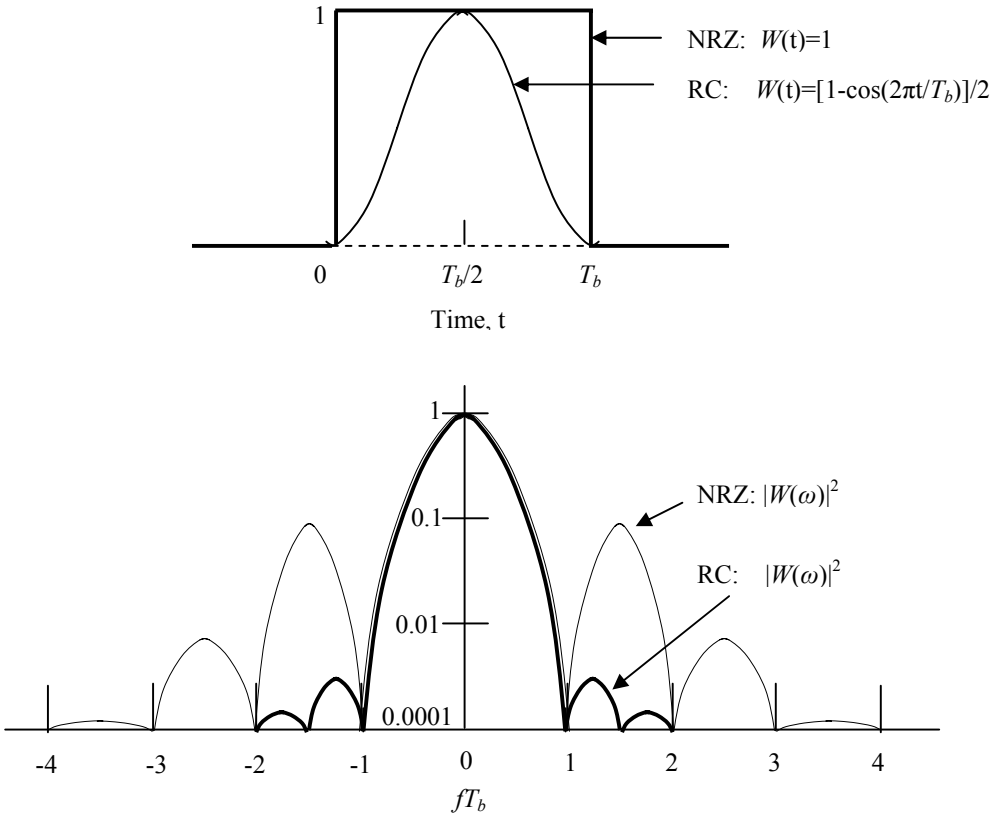


Figure 12.4 NRZ and RC waveform and their spectra

12.2.3 QPSK (Quadrature Phase Shift Keying), OQPSK, MSK

In QPSK, two encoded waveforms $m_c(t)$ and $m_s(t)$ are simultaneously used to BPSK phase modulate the same carrier with quadrature phases. These two encoded waveforms can be either the output from two different digital sources or the output of the alternate bits from a common source. Within one quadrature bit time, the carrier is actually sending two bits. It is therefore that we define

$$T_b = \frac{T_s}{2} , \tag{12.21}$$

where T_b=Bit time,
T_s=Symbol time.

Should these two encoded waveforms be the output of alternate bits from a common source, the transmitted data rate in a QPSK system would be cut by half in contrast to a

BPSK system. The transmitted data rate is proportional to the bandwidth to be occupied by an encoding system. In other words, it means that the frequency throughput in a QPSK system is double of that in a BPSK system. This is the main advantage of a QPSK system in comparison to a BPSK system.

The QPSK carrier can be expressed as

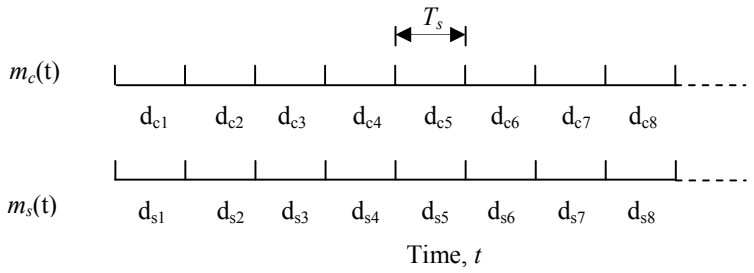
$$C(t) = Am_c(t)\cos(\omega_c t + \Psi) + Am_s(t)\sin(\omega_c t + \Psi) \ . \tag{12.22}$$

Or,

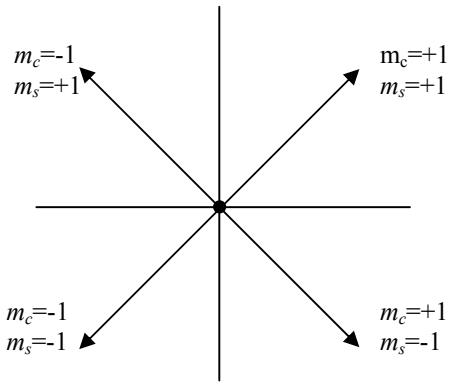
$$C(t) = \sqrt{2}A\alpha(t)\cos(\omega_c t + \theta(t) + \Psi) \ , \tag{12.23}$$

$$\sqrt{2}\alpha(t) = \sqrt{m_c{}^2(t) + m_s{}^2(t)} \ , \tag{12.24}$$

$$\theta(t) = \tan^{-1}\left[m_s(t)/m_c(t)\right] . \tag{12.25}$$



(a) Bit sequence alignment of $m_c(t)$ and $m_s(t)$



(b) Quadrature phase relationship

Figure 12.5 QPSK encoding

Figure 12.5 shows the QPSK encoding scheme. The bit sequences of $m_c(t)$ and $m_s(t)$ are synchronized from each other. These two orthogonal signals can modulate the carrier by

using quadrature amplitude modulation or quadrature phase shifting. Finally, it shows that both $m_c(t)$ and $m_s(t)$ are either +1 or –1 for either NRZ or Manchester waveforms. Thus, the value of $\alpha(t)$ equals to 1 and then QPSK carrier has a constant envelope. The angle $\theta(t)$ is one of the four phases, 45°, 135°, -135°, -45°).

An alternative QPSK is OQPSK (Offset Quadrature Phase Shift Keying). The offset is taken between the $m_c(t)$ and $m_s(t)$ as one-half of a bit time and is shown in Figure 12.6.

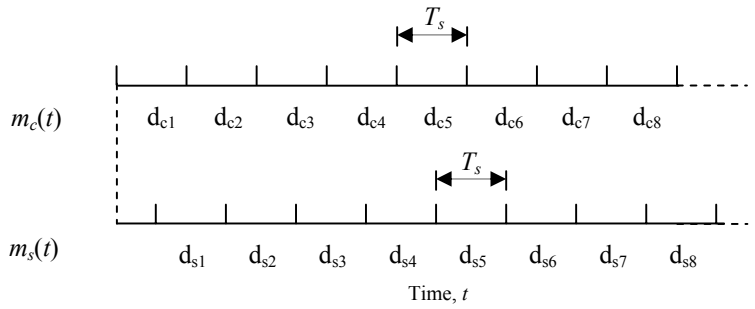


Figure 12.6 Bit sequence alignment of $m_c(t)$ and $m_s(t)$ in OQPSK modulation

The resulting OQPSK carrier can be expressed by

$$C(t) = A\sum_{i=-\infty}^{\infty} d_{ci} W(t - iT_s)\cos(\omega_c t + \Psi) + A\sum_{i=-\infty}^{\infty} d_{si} W(t - \frac{1}{2}T_s - iT_s)\sin(\omega_c t + \Psi) \ . \tag{12.26}$$

The spectrum of OQPSK is the same as that of QPSK as long as these two sequences, $m_c(t)$ and $m_s(t)$, are independent. In contrast with QPSK, OQPSK has two advantages. The offset limits the total phase shift that must be provided at each bit change in QPSK. The offset also provides spectral and interference advantages when the decoding is imperfect or when its non-linearity is acceptable.

An MSK (Minimum Shift Keyed) carrier is a special OQPSK carrier, in which the NRZ waveform is replaced by a half-period sine waveform, that is,

$$W(t) = \sin(\frac{\pi}{T_s})t \ , \qquad \text{when } 0 \le t \le T_s \ , \tag{12.27}$$

and

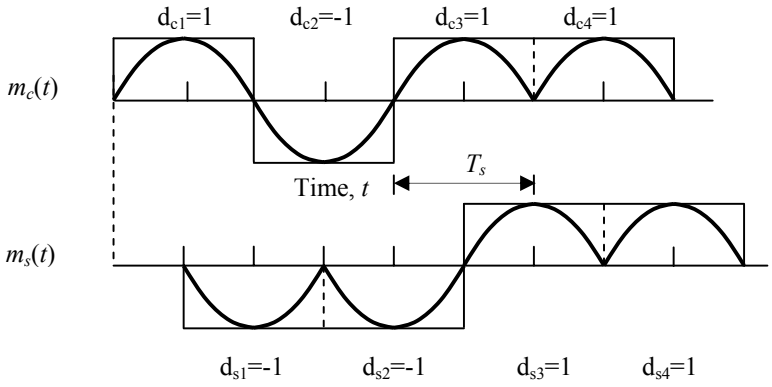
$$W(t - \frac{1}{2}T_s) = \sin\left[\frac{\pi}{T_s}(t - \frac{1}{2}T_s)\right] = \cos(\frac{\pi}{T_s})t \ , \ \text{when } 0 \le t \le T_s \ . \tag{12.28}$$

Substitute (12.27) and (12.28) into (12.26), the resulting MSK carrier is

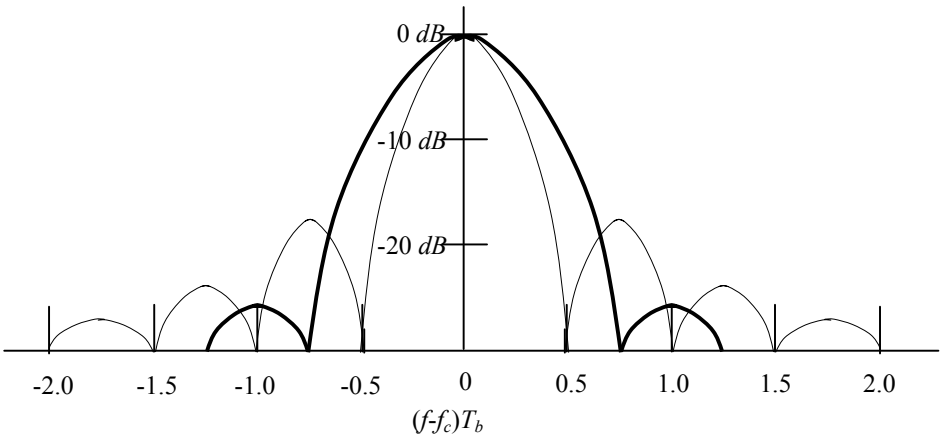
$$C(t) = A\sum_{i=-\infty}^{\infty} d_{ci} \sin\left[\frac{\pi}{T_s}(t - iT_s)\right]\cos(\omega_c t + \Psi) + A\sum_{i=-\infty}^{\infty} d_{si} \cos\left[\frac{\pi}{T_s}(t - iT_s)\right]\sin(\omega_c t + \Psi) \ . \tag{12.29}$$

Finally, by means of triangular formula, it can be derived to

$$C(t) = A\cos\left[\omega_c t + d_i \frac{\pi}{T_s}(t - iT_s) + \chi\right], \qquad \text{when } iT_s \le t \le (i+1)T_s, \tag{12.30}$$



(a) Baseband waveform — MSK, OQPSK



(b) Carrier spectrum — MSK, OQPSK

Figure 12.7 MSK and OQPSK baseband waveform and spectrum

From expression (12.21), we have

$$C(t) = A\cos\left[2\pi(f_c + \frac{d_i}{4T_b})t - id_i\pi + \chi\right], \quad \text{when } i2T_b \le t \le (i+1)2T_b, \quad (12.31)$$

where

$$d_i = d_{ci}, \quad \text{when } iT_b \le t \le (i+1)T_b, \quad (12.32)$$

$$d_i = d_{si}, \quad \text{when } (i+1)T_b \le t \le (i+2)T_b, \quad (12.33)$$

$$\chi_i = \chi_{i-1} + \frac{\pi i}{2}(d_{i-1} - d_i) + \Psi. \quad (12.34)$$

Expression (12.31) indicates a constant envelope carrier. This is one of the main features of the MSK carrier. Also, according to the data bits, it is a carrier with a frequency shift between (f_c-1/(4T_b)) and (f_c+1/(4T_b)). The tone spacing in an MSK is one-half that employed for non-coherently demodulated orthogonal FSK. It is therefore named as "minimum shift keying."

Figure 12.7 shows a MSK baseband waveform and its carrier spectrum in reference to OQPSK. The spectral tail in MSK modulation is significantly reduced from that of OQPSK modulation. This is due to the smooth sinusoidal waveform in MSK modulation in contrast to the rectangular and pulsed waveform in OQPSK modulation. However, a MSK signal requires bandwidth expanded about 50% more than an OQPSK signal.

12.2.4 FSK (Frequency Shift Keying), CPFSK

In an FSK (Frequency Shift Keying) modulation, the NRZ bit waveform modulates the carrier so that the carrier frequency is shifted between two modulation frequencies according to the bit waveform. The encoding is therefore referred to as frequency shift keying. The modulated carrier has a constant envelope and the form

$$C(t) = A\cos\left[\omega_c t + 2\pi\Delta_f \int m(t)dt + \Psi\right], \quad (12.35)$$

where $m(t)$ = the bit sequence.
It is somewhat difficult to compute the carrier spectrum. Its main hump occupies an area from -0.5(f-f_c)T_b to +0.5(f-f_c)T_b.

If the NRZ waveform is replaced by an RC waveform, the frequency-shifted carriers have the property of continuous phase change. Thus, the modulated carrier can be decoded by phase demodulation, which has decoding advantages over the standard frequency detection. This type of FSK is often referred to as CPFSK (Continuous Phase Frequency Shift Keying).

12.3 Decoding and Bit-Error probability

In analog communication systems, the main concern is to preserve signal waveforms by producing an *SNR* ratio as high as possible while in digital communication systems, the main concern is to recover digital bits by reducing BER (Bit Error Rate) as low as possible.

The signal waveforms in digital communication are converted to sequences of data bits or symbols, which are denoted as "0" and "1". Bit error means that a "1" is decoded as "0", or vice versa. The probability of making a bit error had been computed for various decoding schemes and is listed in Table 12.3.

Table 12.3 BER (Bit Error Rate) in various decoding schemes

Decoding scheme	BER (Bit Error Rate)
BPSK, QPSK, MSK	$Q[\ (2E_b/N_o)^{0.5}]$
Coherent FSK	$Q[\ (E_b/N_o)^{0.5}]$
DPSK	$e^{-Eb/No}/2$
Noncoherent FSK	$e^{-Eb/2No}/2$

From Table 12.3 it can be seen that the BER is a function of only the ratio E_b/N_o, for all the decoding schemes, that is,

$$BER = f\left(\frac{E_b}{N_o}\right), \tag{12.36}$$

where

$$Q[x] = \frac{1}{2\pi}\int_x^\infty e^{-t^2/2}dt\ , \tag{12.37}$$

$$\frac{E_b}{N_o} = \frac{Bit\_energy\_of\_the\_modulated\_carrier\_at\_decoder\_input}{Spectral\_level\_of\_additive\_noise\_at\_decoder\_input}, \tag{12.38}$$

or,

$$\frac{E_b}{N_o} = \frac{P_c T_b}{N_o} = \frac{P_c}{N_o R_b}\ , \tag{12.39}$$

where E_b = Bit energy,
P_c = Power of carrier,
T_b = Bit period,
N_o = Noise power spectral density,
R_b = Bit rate.

Expression (12.36) is an extremely significant result in decoding theory. That is, the BER depends on only the ratio of the bit energy to the noise power. This fact greatly simplifies bit error analysis.

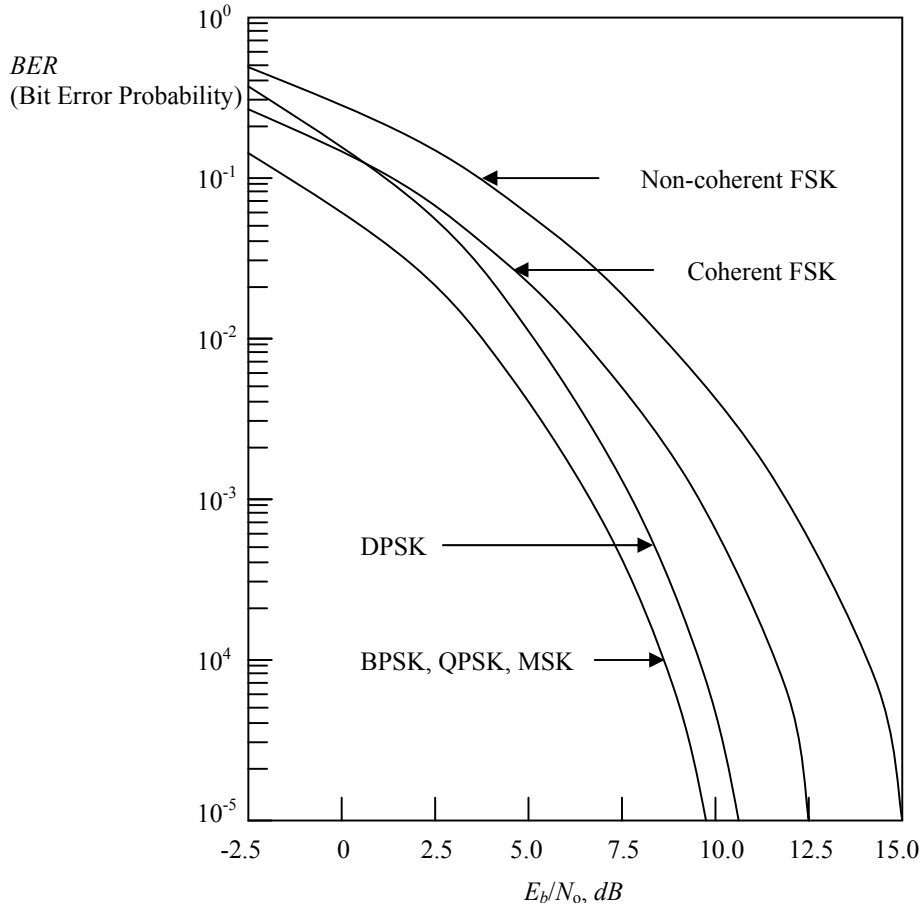


Figure 12.8 Relationship between BER and E_b/N_o.

It should be noted that there are two decoding operations, coherent and non-coherent. Coherent decoding requires the decoder to use a referenced carrier at the same frequency and phase as the received modulated carrier during each bit period. A synchronization subsystem is therefore required in the receiver. On the contrary, non-coherent decoding does not require a reference carrier. For example, in the differential PSK(DPSK) or non-coherent PSK encoding, a BPSK carrier is used. The carrier phase of each bit is referenced to the previous bit. A binary, +1 or –1, is encoded by using either the same phase or a π shift of the previous bit phase, depending on whether the present bit is identical or opposite to the previous bit. It follows the rule of

$$d_i = d_{i-1} a_i \ . \tag{12.40}$$

where a_i = present bit.

Based on the expressions in Table 12.2, the curves to show the relationship between BER and the ratio, E_b/N_o, for various encoding/decoding schemes are plotted in Figure 12.8.

As shown in Figure 12.8, the bit error rate is reduced as the ratio E_b/N_o is increased. For example, in order to reduce the BER lower than 10^{-5}, the ratio must be increased over 10 dB for all the modulation schemes. Among the various modulation schemes, BPSK, QPSK, and MSK have the same (and lowest) BER. Also, it can be seen that coherent modulation always has lower BER than non-coherent modulation. The stipulation is that more effort must be put for coherent modulation.

12.4 Error Correction Schemes

In digital communication, it is the noise in a communication channel that causes bit error in the decoding. Figure 12.8 shows that the higher the ratio E_b/N_o, the lower the bit error probability. From expression (12.39), it is obvious that the ratio E_b/N_o is proportional to the carrier power P_c and counter-proportional to the bit rate R_b. It implies that under the same value of carrier power P_c the higher the bit rate R_b to be transmitted, the lower the ratio E_b/N_o, consequently, the higher the bit error probability.

Intuitively the bit error probability in the decoding could be reduced by 50% if a bit is transmitted twice with the same bit energy. As a result, the effective bit rate or data rate will be slowed down 100% due to the twice transmission for a bit. However, in order to maintain the same data rate or the effective bit rate, the bit rate must be increased to two times the original bit rate. It results in a 50% decrease of the bit energy if the carrier power is maintained as the same before and after the bit rate is increased. The decrease of bit energy eventually pushes the bit error rate up. Now, is the resulting bit error probability reduced or increased? As a matter of fact, it is hard to reduce bit error probability using such a simple technology and in reality, it somehow increases the bit error probability.

Instead of the simple scheme of transmitting the data bits twice, other error correction schemes have been developed in the digital communication systems. They could basically be categorized into two groups: block encoding and convolutional encoding. The common idea is to add redundant bits to data bits so that the bit error in the decoding could be corrected by the combination of redundant and data bits through an appropriate algorithm.

In error-correction block encoding, each block of k data bits is first encoded to a new block with n bits, in which $(n-k)$ bits are redundant for the error correction. Of course,

$$n > k \ . \tag{12.41}$$

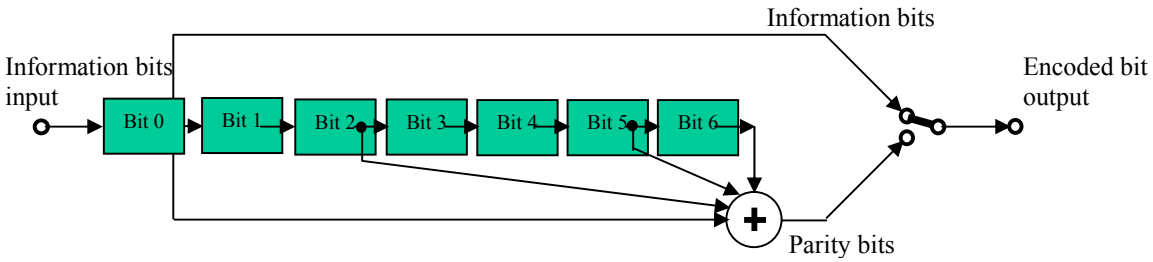
And, the code rate, r, is defined as

$$r = \frac{k}{n} \ . \tag{12.42}$$

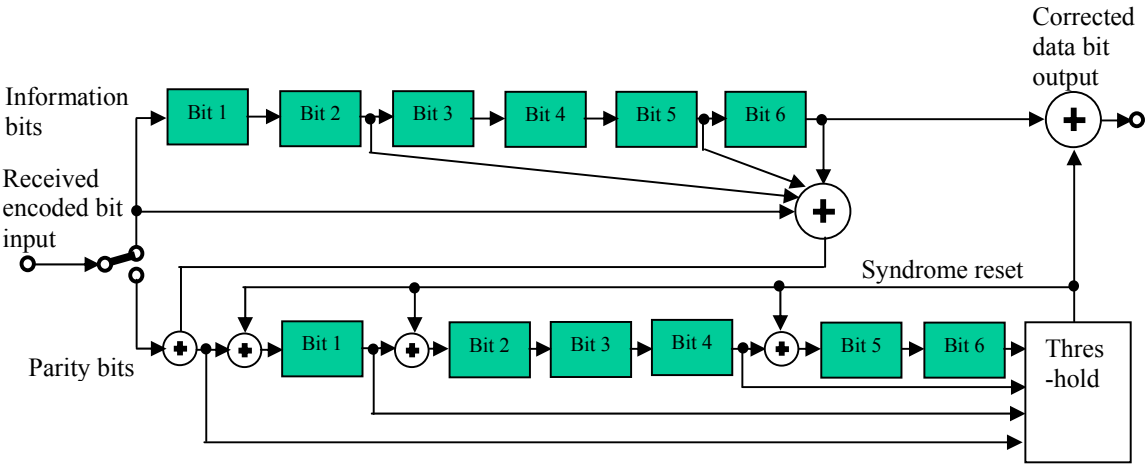
According to the prescribed coding rule the redundant bits are encoded. One popular type of redundant bit is the parity-type bit, in which the redundant bit is determined on the basis of the parity of the previous or subsequent bit. If a bit is decoded incorrectly, it would be corrected from the parity bit based on the prescribed coding rule.

In convolutional encoding, rather than a fixed data block, the data bits are interleaved with parity bits while parity bits are produced with prescribed rules. In addition, some other redundant bits, such as the CRC (Cyclic Redundancy Check) bits are added as additional information bits so that bit error could be more effectively corrected. In the

process of decoding, the data bits are decoded and corrected by the parity and CRC bits on the basis of the prescribed coding rules in the encoder.



(a) Encoder



(b) Decoder

Figure 12.9 A convolutional coding plan, r=1/2, k=7

Figure 12.9 shows an example of a convolutional coding for error correction. Either encoder or decoder is basically formed by using shift registers. There is one AND block in the encoder while there are four AND blocks in the decoder. The threshold block in the decoder executes the error correction for the information bits. The information bits are the primary data bits while the parity bits are the redundant bits. The code rate, r=1/2, shows that the number of the parity bits equals to that of the information bits. The integer, k=7, is the constraint length. It represents the number of shift registers in the encoder.

The advantages of block coding are its simplicity and its independence from block to block. On the contrary, the advantages of convolutional coding are its shorter word-lengths, special protection ability against burst errors, and the lower word error probability.

Error correction is a big subject in digital communications. The purpose of this section is just to introduce the fundamental understanding about the communication systems for an RF designer. Obviously, detailed discussion is beyond our scope and can be found in specific published communication books or papers. As a matter of fact, in addition to error correction, some other topics, such as the mode of multiple access, FDMA (Frequency Division Multiple Access), TDMA (Time Division Multiple Access), CDMA (Code Division Multiple Access), and so on, are very important in communication systems.

References

[1] R. E. Ziemer and W. H. Tranter, “Principles of Communications,” (Book), Houghton Mifflin Company, 1976.

[2] S. Pasupathy, “Minimum Shift Keying: A Spectrally Efficient Modulation,” IEEE Communications Magazine, pp. 14 -22, July 1979.

[3] K. Murota et al., “GMSK Modulation for Digital Radio Telephony,” IEEE Transactions on Communications, Vol. COM-29, No. 7, pp. 1044 -1050, July 1981.

[4] B. P. Lathi, “Modern Digital and Analog Communication Systems,”(Book), Holt, Rinchart, and Winston, New York, NY 1983.

[5] Robert M. Gagliardi, “Satellite Communications,” (Book), Lifetime Learning Publications, 1984.

[6] Jack Smith, “Modern Communication Circuits,” (Book), McGraw-Hill Publishing Company, 1986

[7] Bernard Sklar, “Digital Communications,” (Book), Prentice Hall, 1988.

[8] J. Socci, “GMSK As A Modulation Scheme for A Frequency Hopped PHY,” IEEE 802.11 Wireless Access Method and Physical Layer Specifications, July 1993.

[9] B. Madsen, and D. E. Fague, “Radios For The Future: Designing for DECT,” RF Design, pp. 48-53, April 1993.

[10] D. E. Fague, “A Fully Integrated Modulator/Demodulator for DECT,” RF Expo West, San Jose, CA, March 17-19, 1993.

[11] B. Wuppermann et al., “A 16-PSK Modulator With Phase Error Corrections,” ISSCC93, San Francisco, CA, February 1993.

[12] A. Boveda et al., “A 07-3 GHz GaAs QPSK/QAM Direct Modulator,” ISSCC93, San Francisco, CA, February 1993.

[13] A. Boveda et al., “A 07-3 GHz GaAs QPSK/QAM Direct Modulator,” IEEE Journal of Solid State Circuits, Vol. 28, No. 12, December 1993.

[14] D. L. Ash, “SAW devices in wireless communication systems,” Ultrasonics Symposium, 1993. Proceedings., IEEE 1993 , Vol. 1, pp. 115 – 124, 31 Octocber -3 November 1993.

[15] K. Negus et al., “Highly integrated transmitter RFIC with monolithic narrowband tuning for digital cellular handsets,” 1994 IEEE International Solid-state Circuits Conference, pp.38-39, February 1994.

[16] K. Pahlavan, and A. H. Levesque, “Wireless Data Communications,” Proceedings of IEEE, Vol. 82, No. 9, pp.1398-1430, September 1994.

[17] A. Flatman, “Wireless LANs: Developments in Technology and Standard,” IEEE Computing & Control Engineering Journal, Vol. 5, No. 5, pp.219-224, Octocber 1994.

[18] T. Stetzler et al., “A 2.7V to 4.5V Single Chip GSM Transceiver RF Integrated Circuit,” 1995 IEEE International Solid-state Circuits Conference, pp. 150-151, February 1995.

[19] J. E. Padgett, C. G. Gunther, and T. Hattori, “Overview of Wireless Personal Communications,” IEEE Communications Mag., Vol. 33, No. 1, pp. 28-41, January 1995.

[20] P. K. Halder, K. B. Letaief, “Performance analysis of spread-spectrum multiple-access with DPSK modulation in indoor wireless communications systems,” Universal Personal Communications. 1995. Record., 1995 Fourth IEEE International Conference on , pp. 828 – 832, 6-10 November 1995.

[21] Min-Lu Dai, “Planning indoor wireless communication system,” Personal, Indoor and Mobile Radio Communications, 1998. The Ninth IEEE International Symposium on , Volume: 1 , Vol. 1, pp. 74 – 78, 8-11 September 1998.

[22] Y. Matsumoto, N. Mochizuki, M. Umehira, “OFDM subchannel space-combining transmission diversity for broadband wireless communication systems,” Universal Personal Communications, 1998. ICUPC '98. IEEE 1998 International Conference on , Volume: 1 , Vol. 1, pp. 137 – 141, 5-9 Oct. 1998.

[23] J. Singh, “Mixed-signal data conversion chip for wireless communication systems,” Wireless Communications and Networking Conference, 1999. WCNC. 1999 IEEE , Vol. 2, pp. 737 – 741, 21-24 September 1999.

[24] Ruly Lai-U Choi; K.B.Letaief, R.D.Murch, “Transmit diversity combining for wideband DS/CDMA in wireless communication systems,” Computers and Communications, 2000. Proceedings. ISCC 2000. Fifth IEEE Symposium on , pp. 717 – 722, 3-6 July 2000.

[25] S. J. Spiegel, “A broadband receiver architecture for wireless communication systems,” Electrical and ELectronic Engineers in Israel, 2000. The 21st IEEE Convention of the , pp. 121 – 124, 11-12 April 2000.

[26] T. Yazaki, H. Yamamoto, A. Hyogo, K. Sekine, “Low-power ASK receiver circuit for wireless communication system,” Circuits and Systems, 2002. APCCAS '02. 2002 Asia-Pacific Conference on , Vol. 2 , pp. 405 – 408, 28-31 Oct. 2002.

[27] H. Y. D. Yang, “Analysis of RF radiation interference on wireless communication systems,” Antennas and Wireless Propagation Letters , Vol. 2 , Issue: 9 , pp. 126 – 129, 2003.

[28] L. Angrisani, M. D'Apuzzo, M. D'Arco, “ A new method for power measurements in digital wireless communication systems,”Instrumentation and Measurement, IEEE Transactions on , Volume: 52 , Issue: 4 , pp. 1097 – 1106, August 2003.

Index