

# Analog

*Inexact Science, Vibrant Art*

(PRELIMINARY DRAFT: Highly Incomplete!)

ALI HAJIMIRI

*Caltech*

DRAFT 1/2021

Copyright ©1998–2021 Ali Hajimiri

Permission is granted for personal and individual use of this document. You may print one copy for your personal use. Permission is **NOT** given to distribute, place this document on a secondary website, or print more than one copy of this document. Permission is **NOT** given to monetize this document in any shape or form. **Do NOT distribute this copy.** If somebody asks you for a copy please send them to the following web page to download the latest version.

This is a living document. The latest version can be downloaded from:  
[Caltech Holistic Integrated Circuits \(CHIC\) Resources Page.](#)

*To the memory of my father,*

*Dr. Javad Hajimiri,*

*A man ahead of his time,*

*in life,*

*and,*

*in death.*



Photo Credit: Emad Hajimiri

DRAFT 1/2021

# Preface

When I started as a new assistant professor at Caltech in 1998, I began teaching a two-quarter class on analog circuit design (EE114) (which is still being offered to this date). I thought there was not sufficient focus on the fundamentals and the underlying thought process that goes into circuit design. In my view, approaches fell into one (or a combination) of several categories. Some focused on “exact analysis” based on symbolic and algebraic analysis, where the essence of the situation was almost lost and had to be excavated from underneath a pile of terms using various approximations. On the other end of the spectrum were some others that attempted to provide intuition at the cost of over-simplifications that sometimes led to clearly incorrect results when applied to other situations. Yet, some others made an attempt at cataloging as many different circuit topologies as possible.

I was hoping to convey to the class my impression of the underlying themes and underpinning principles that go into analog circuits. As I started getting ready to teach, I frantically compiled my own handouts for the class, instead of relying on any given textbook. However, after a while, the research and other responsibilities stopped me from continuing to type my handouts. Over the years, whenever I came up with a somewhat new way to look at the subject matter and found a little respite “on the thin ice of modern life,” I added an additional chapter. I had a much grander vision of what this “book” will eventually include and waited till I had time to complete it.

“Then one day [I found], ten<sup>1</sup> years have got behind [me]” and that I “missed the starting gun.” So now, I “run and [I] run to catch up with the sun” by making what I have available to everyone, **as is**. This is a **highly incomplete document** with lot of holes, errors, typos, poorly drawn (or outright missing) figures, and notes to self. It is missing more than half of the intended content and does not even come close to the editorial standards of any of the textbook on this subject matter out there. But I cannot wait any longer, so here is my “green leaf”<sup>2</sup> to you, dear reader. If the universe permits me, I still intend to update this manuscript over time and complete it<sup>3</sup> at some point. This material goes hand-in-hand with the [analog circuit design lectures](#) on YouTube that

<sup>1</sup>Actually more like fifteen, but who is counting.

<sup>2</sup>The old Persian proverb roughly translates as: “A green leaf is Dervish’s offering, as that’s all he has to offer.”

<sup>3</sup>In my view, in its current form it is less than one third complete.

go farther than this manuscript. There are also a fair number of problems designed to go with this material that might be released at a later time.

## 0.1 Philosophy

In circuit design, the objective of analysis should be to use the least complicated tools that capture the essence of problem, without over-simplifications that can result in inaccurate and potentially misleading results. The objective of analysis should be to provide the designer with a clear understanding of the causal relation among different design parameters, topological choices, and the performance. In this manuscript, I discuss this mindset and present some of these tools and principles. My goal is to provide the reader with several analysis tools with varying degrees of accuracy and ease of use.

“*Exact analysis*” in the most general sense is nothing but a myth. To arrive at the equivalent models for circuit analysis, we have gone through numerous layers of approximations within different levels of abstraction. To recap, the entire circuit analysis is based on KCL and KVL that are special cases of derived from Maxwell equations for lumped circuits when the dimensions are much smaller than the wavelength. The transistor models are based on numerous simplifying assumptions about the dominant physical phenomena, as discussed in Chapter 1. Even beyond that we make another great leap whenever we use the linearized small-signal models to analyze devices that are fundamentally, and often substantially, nonlinear in nature. So spending a fair amount of time performing an “*accurate analysis*” of a circuit, let us say using nodal analysis has to be well justified before being performed. Considering the strength and broad availability of circuit simulators, doing an exact (or even approximate) analysis of a circuit *just* to predict the result is an exercise in futility most of the times.

Nonetheless, the above arguments simply do *not* mean that there is no value to hand analysis and quite on the contrary would mean that we need to be even more proficient in analysis to be effective in designing circuits. The hand-analysis is an extremely power tool that comes to the aid of a good designer to understand different trade-offs involved in the design of a circuit and get a sense for the relative strength of various effects in a circuit. In general, good hand analysis is not a substitute for computer simulation, rather it should be complimentary to computer simulations. Circuit simulators are good at telling you if something does not work, but often clueless as to what needs to be done to make it work. Design oriented analysis does not just provide the designer with *quantitative* (and usually incremental) improvements to the circuit that computer simulators are so good at. It is real strength is in providing the designer with a beacon toward *qualitative* changes to the circuit (e.g., topology changes) that can result in significant non-incremental improvements.

The approximate nature of any circuit analysis has another important and somewhat counter-intuitive implication about what kind of analysis tools we need to develop. We already know that we can obtain the most accurate result

within the circuit level of abstraction using classical approaches (e.g., nodal analysis for linear circuits) and even better using computer simulators. The most accurate analytical tools will be as accurate as the nodal analysis approach.

To be able to understand the appropriate analytical tool, we have to have a deep understanding of the capabilities and limitations of each tool. While probably 80% of problems in day-to-day design can be dealt with using the simplest analysis techniques, these are often considered trivial and quickly solved. On the other hand, it is the remaining 20% of challenges that take 80% (and sometimes a lot more) of the designers time. That is why it is not just an academic exercise to spend a lot of time considering more interesting (read complicated) and sometimes obscure (read pathological) cases. These cases may not happen all the time, but when they happen they take a lot time to understand and resolve and have the added side benefit of separating the wheat from the chaffOK, that is a little bit over the top, but then again, I have always had a weak spot for the dramatic!. It can be seen that the most effective creator is the one who knows all his tools very well so he can use the appropriate one when needed. Good design oriented analysis does not just (over-)simplify every problem, but it uses the simplest tool appropriate for solving that particular problem.

The salient feature of design oriented tools is their modular nature that allows them to be applied in a progressive fashion, with each step providing additional information about the problem and providing a more accurate approximation of the problem. This allows one to arrive at a first order understanding of the problem without having to carry out the analysis to the (sometimes bitter) end and having to make simplifying assumptions then. More accurate results can be obtained by successive application of the method. Nodal analysis (i.e., writing KCL and KVL in an organized fashion) is a perfect example of an approach not fitting this criterion. One needs to carry to complete analysis to the end to find a solution that often needs to be simplified. This is not to say that nodal analysis is not useful. On the contrary it is a very powerful tool to write simulators and to understand the limitations of various analytical approaches and why and when they fail under certain circumstance.

The other important feature of design-oriented analysis techniques is their ability to identify the dominant limiting effects in the circuit, where the design effort should be focused. Everyone agrees with the logical argument that in dealing with the technical problem one should focus on the dominant, first-order effects first before focusing on less-dominant (and often cosmetic changes). Nevertheless, most of us have fallen victim to spending a lot of time solving a non-dominant problem. This is where a step-wise modular approach is most appropriate.

In the subsequent sections we introduce some of these tools. As learning is an inductive (top-down) process and recalling is often a deductive (bottom-up) one, we discuss these approaches in an ascending order of complexity and accuracy. With discussion about when they are appropriate to use. It should be no surprise that the last tool in the toolbox is the most accurate and general and subsumes the other, but the learning process being an inductive one, we start with the simplest and generalize.

## 0.2 Acknowledgements

Nothing is done in isolation. I would like to thank Dr. Brian Hong and Dr. Abbas Komijani for having made important contribution to this work. Also, I am grateful to Parham Porsandeh Khial, Austin Fikes, Richard Ohanian, and Craig Ives for doing a professional caliber job of recording the lectures over the years. I am greatly indebted to a large group of people including Dr. Brian Hong, Mr. Austin Fike, Mr. Parham Porsandeh Khial, Mr. Matan Gal-Katziri, Mrs. Carol Sosnowski, Mr. Richard Ohanian, Mr. Reza Fatemi, Mr. Craig Ives, Mr. Aroutin Khachaturian, Mr. Stefan Turkowski, Dr. Kaushik Dasgupta, Dr. Alex Pai, Dr. Amirreza Safaripour, Mr. Tomoyuki Arai, Ms. Jennifer Arroyo, Prof. Aydin Babakhani, Dr. Florian Bohn, Prof. Steven Bowers, Ms. Jay Chen, Dr. Edward Keehr, Dr. Abbas Komijani, Dr. Behnam Analui, Dr. Shohei Kosai, Prof. Arun Natarajan, Prof. Kaushik Sengupta, Prof. Costis Sidiris, Prof. Hua Wang, Dr. Behrooz Abiri, Prof. Yu-Jiu Wang, Mr. Alexander White, and Mr. J. Yoo formerly or currently of Caltech as well as Prof. Boris Murmann of Stanford University and Prof. Hossein Hashemi of USC for their valuable feedback on the manuscript. I am sure I am missing somebody and if you notice one, please let me know.

I am very grateful to my family for always being supportive of me. My parents, as well as my wife and my children have been the shining light of my life.

I was inspired by late R. David Middlebrook, who taught a very different course on circuits with the same number (EE114) at Caltech before me. His design-oriented analysis approach seemed like the right mindset. Although the generalized TTC method and the subsequent discussions on feedback are not the same as the extra-element theorem method that Middlebrook developed, they were inspired by his mindset, which seemed to find a useful (and hopefully happy) medium between brute-force algebra and watered-down intuition.

Last, but not least, I am particularly indebted to the students in Caltech's EE44 and EE114 series who helped me develop and improve this material over more than two decades by asking great questions and refusing to accept incomplete answers.

*Ali Hajimiri  
April 2020 (COVID-19 lock down)*

# Chapter 1

## Underlying Physics

### 1.1 Semiconductor Materials

In this section we discuss some of the underlying principles of semiconductor materials.

#### 1.1.1 Macroscopic Properties

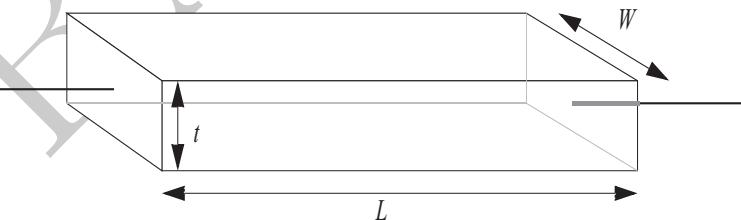
Online YouTube lectures:

[\*\*101N. Basic Solid-State Physics: Energy bands, Electrons and Holes\*\*](#)

The resistance of a slab of an arbitrary material is proportional to its length and inversely proportional to its cross-section area (Figure 1.1), namely,

$$R = \rho \cdot \frac{L}{A} = \rho \cdot \frac{L}{t \cdot W} \quad (1.1)$$

The proportionality constant is called resistivity of the material and is a characteristic of the material and independent of its shape. When the thickness of



$$R = \rho \frac{L}{A} \qquad A = Wt$$

Figure 1.1: A slab of resistive material

the layer is fixed (*e.g.*, in integrated circuits, it is more convenient to write (1.1) as

$$R = R_{\square} \cdot \frac{L}{W} \quad (1.2)$$

where  $R_{\square} = \rho/t$  is known as sheet resistance. This allows the value of the resistor to be controlled by adjusting the length-to-width ratio (often measured in number of squares).

Resistivity,  $\rho$ , is measured in units of ohm-meter [ $\Omega \cdot m$ ] and varies over many orders of magnitudes. For example the resistivity of aluminum is on the order of  $10^{-8}\Omega \cdot m$ , while the resistivity of silicon-dioxide ( $SiO_2$ ) is on the order of  $10^{14}\Omega \cdot m$ . Note that there are about 22 orders of magnitude difference between the resistivity of aluminum and  $SiO_2$ . Interestingly, both of these materials are used extensively in integrated circuits. Different materials can be divided into three categories: insulators, semiconductors, and conductors. This categorization can be done based on the resistivity of the material, as shown in Figure 1.2.

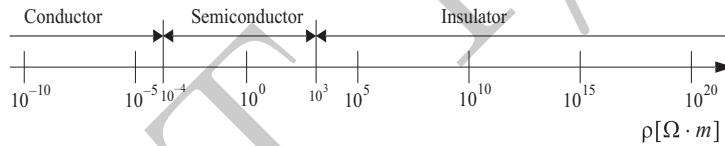


Figure 1.2: The resistivity range of materials

### 1.1.2 Microscopic View

Conduction properties of materials can be better understood by looking at the microscopic structure of the materials. We will utilize the simplest model that explains the properties of semiconductor material of interest to us. Remember that a particle having a momentum of  $p$  has a De Broglie wavelength given by

$$\lambda = \frac{h}{p} \quad (1.3)$$

where  $h$  is the Planck's constant. Let us look at the case of a hydrogen atom, symbolically shown in Figure 1.3. For the electron to have a stable orbit, the wavelength has to form 'standing De Broglie waves', as shown in Figure 1.4. The condition for formation of such standing wave is for the circumference to be an integer multiple of the wavelength, i.e.,

$$2\pi r = n\lambda = \frac{nh}{p} \quad (1.4)$$

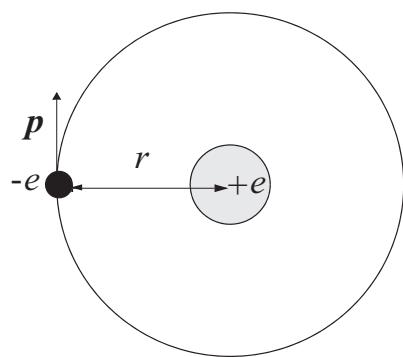


Figure 1.3: Bohr's model for Hydrogen atom

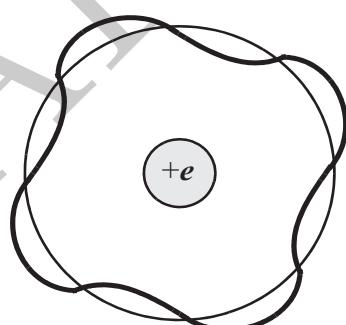


Figure 1.4: Electron's standing wave in the orbit

with  $r$  being the radius and  $n$  an integer. The electrostatic attraction between the electron and the nucleus is given by Coulomb law:

$$F = \frac{1}{4\pi\epsilon_0} \cdot \frac{e^2}{r^2} \quad (1.5)$$

Using Newton's second law, we can write:

$$\frac{1}{4\pi\epsilon_0} \cdot \frac{e^2}{r^2} = ma = m \frac{v^2}{r} = \frac{p^2}{mr} \quad (1.6)$$

Using the condition for standing waves and the above equation we have two equations to solve for  $p$  and  $r$ . Doing so we obtain (Bohr's Hydrogen model):

$$\begin{aligned} p &= \frac{m_0}{2\epsilon_0} \cdot \frac{e^2}{nh} \\ r &= \frac{n^2 h^2 \epsilon_0}{\pi m_0 e^2} \end{aligned} \quad (1.7)$$

The total energy is the sum of kinetic and potential energy:

$$E_n = E_p + E_k = -\frac{1}{4\pi\epsilon_0} \cdot \frac{e^2}{r} + \frac{mv^2}{2} = -\frac{p^2}{2m} = -\frac{m_0 e^4}{8\epsilon_0^2 h^2} \cdot \frac{1}{n^2} = -\frac{13.6 eV}{n^2} \quad (1.8)$$

Therefore, for negative total energy of electrons, there are only certain values of energy an electron can have. Note that the energy band will be continuous for positive energies, i.e. unbounded electron. This is shown in Figure 1.5.

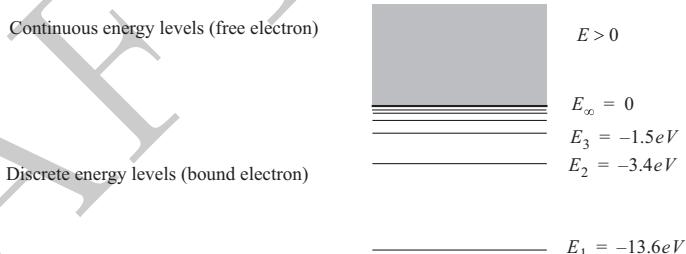


Figure 1.5: Electron's energy levels

When two atoms are far apart, their electrons can have similar energy levels since the electrons have different spacial properties, as pictorially shown in Figure 1.6. However, once the nuclei are brought close to each other, the electrons cannot occupy exactly the same energy bands due to Pauli's exclusion principle and therefore the energy bands will split and become degenerate, as shown in Figure 1.7. This property can be traced back to Pauli's exclusion principle.

Clearly as more nuclei come together to form the crystal lattice, the energy levels split to a larger number of closely spaced energy levels. For a practical piece of material, the number of energy levels becomes so large that they

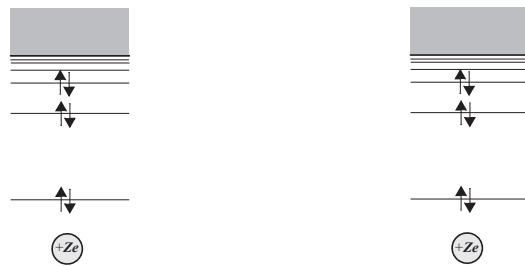


Figure 1.6: Energy levels of two Hydrogen atoms far apart.

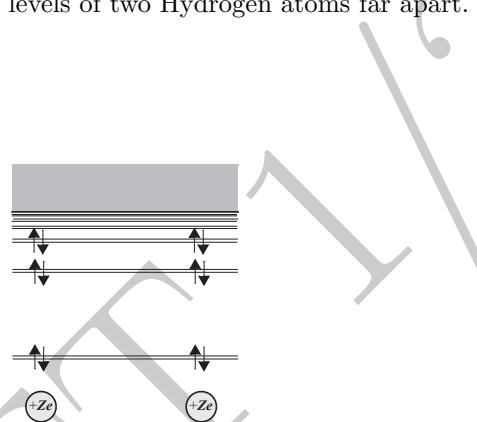


Figure 1.7: Energy levels of two Hydrogen atoms close to each other.

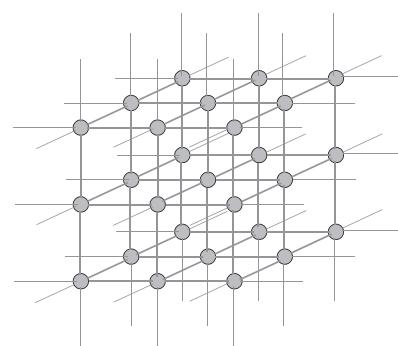


Figure 1.8: Energy levels of a lattice.

practically form 'energy bands' separated by forbidden energy levels, known as 'bandgaps', as shown in Figure 1.8. This effect can be better understood by looking at the energy band structure of Carbon atoms as they are brought close to each other from infinity. The representative a plot of the energy bands versus atomic separation is shown in Figure 1.9. Column IV elements of the periodic

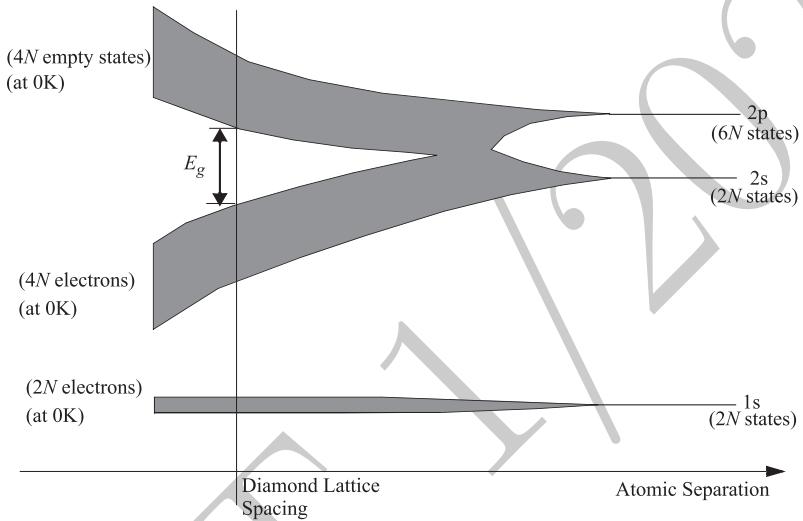


Figure 1.9: Allowed energy levels of a diamond lattice vs. lattice spacing.

table such as Carbon, Silicon and Germanium have a diamond lattice structure. In this structure, each atom shares the 4 electrons in its outmost shell with four adjacent atoms to form covalent bonds. This structure can be shown using the following two dimensional picture of Figure 1.10. Each line represents an elec-

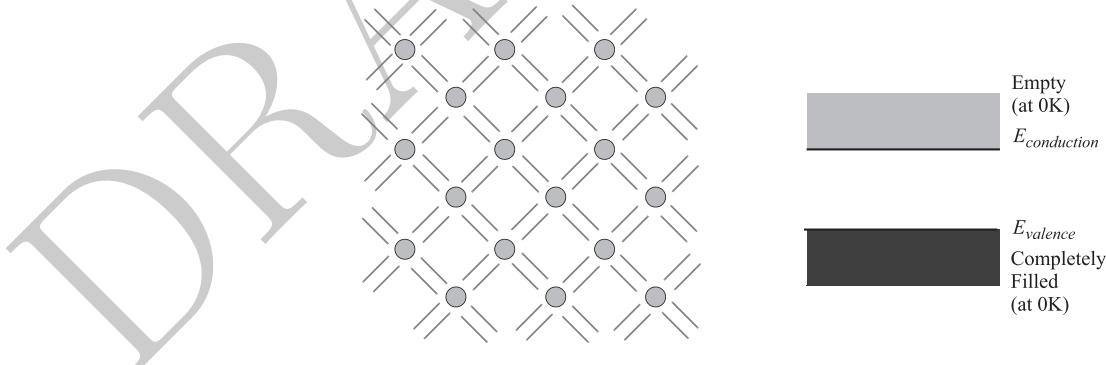


Figure 1.10: 2-dimensional representation of a pure semiconductor at 0K.

tron and two parallel lines represent two electrons with opposite spins forming

a covalent bond. At 0K all the electrons are bonded and therefore no electron can move. The same concept can be understood using the band diagram shown on the right. All the states in the lower energy band (known as valence band) are filled so no single electron can move. Also the higher energy band (known as conduction band) is empty (there is no weakly-bonded electron). The electrons in the valance band cannot form a net current since there are no free states. Also there is no free electron in the conduction band. In fact, pure diamond and silicon can be classified as insulators. A useful analogy is that of two glass cylinders, one completely filled with water and the other one empty. The effect of an external electrical field in the semiconductor is similar to the gravitational force exerted on each molecule of water in the cylinder when it is tilted (as in Figure 1.11). In this case, the net flow of water is zero since there is no space for

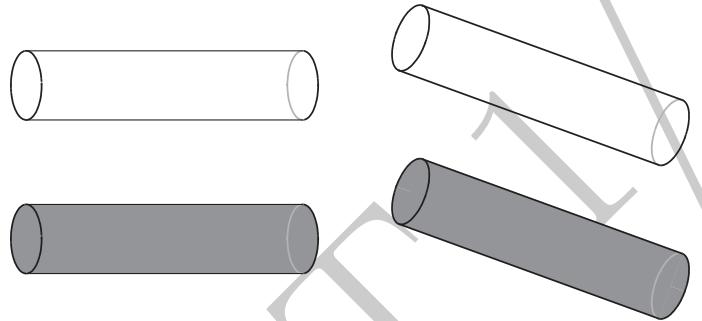


Figure 1.11: Water-filled cylinder analogy for a pure semiconductor at 0K.

it to move to (no empty states). The resistivity of materials can be explained in terms of their band diagrams. Two factors affect the resistivity of materials, namely,

1. The bandgap
2. The number of electrons in each band

Insulators generally have large bandgaps (larger than 4eV) with fully filled valence bands and empty conduction bands. Pure semiconductors are similar to insulator in terms of filled and empty bands but have smaller bandgaps (between 0.5eV to 3eV). Conductors have either overlapping conduction and valence bands (no bandgap) or partially filled conduction bands, as in Figure 1.12).

At temperatures above 0K some electrons have enough thermal energy to cross the bandgap and end up in the conduction band. This process results in an electron in the conduction band and an empty state in the valence band. This process is shown in 1.13.

The electron in the conduction band can move around freely since there are many empty states around it. The 'hole' in the valence band can also move

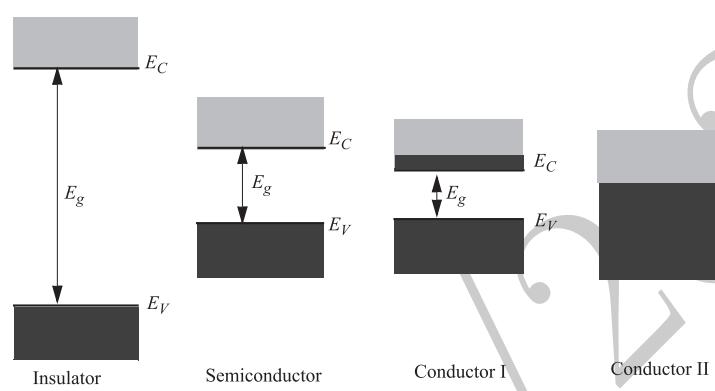


Figure 1.12: Energy band diagrams of an insulator, a semiconductor, and two different kinds of conductors.

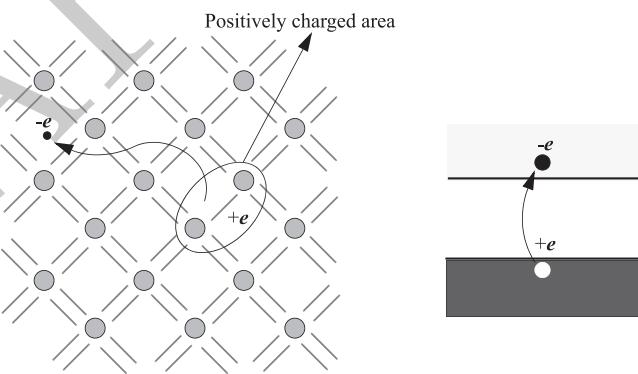


Figure 1.13: Thermal electron-hole generation.

around as electrons from adjacent states fill it and leave a similar 'hole' behind. Note that the area where the hole 'resides' has a positive net charge. Therefore the hole can be viewed as a 'particle' with a charge of  $+e$  in the valance band.

Going back to our water-filled cylinder analogy, if we move a small amount of water from the lower cylinder to the upper one, there will be a bubble in the lower cylinder and a droplet in the upper one. The behavior of the hole-electron pair in the semiconductor material is somewhat similar to the movement of the bubble-droplet pair in the water filled cylinders (Figure 1.14).

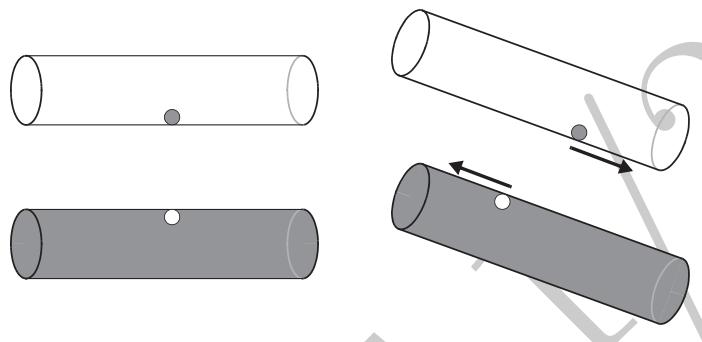


Figure 1.14: Water filled cylinder analogy for thermal generated electron-hole pair.

There are two ways to track the movement of the bubble in the lower cylinder. One way is to track the movements of all the remaining molecules of water. However, it is easier to think of it as a particle with negative mass experiencing the viscosity of the liquid and solving the equations of motion for such a particle. In the same fashion, it is easier to track a hole rather than all the electrons in the valance band. Thus, from now on we will treat holes as real particles with positive charge  $+e$  in the valance band.

It is possible for a free electron in the conduction band to meet another hole somewhere in the semiconductor and jump down to the valence band. When this happens, both electron and hole disappear. This is the reverse process of electron-hole generation and is called recombination. In steady-state, and in the absence of any external effect, the recombination rate is equal to the generation rate. (Otherwise the number of free carriers keeps increasing or decreasing hence violating the definition of steady-state).

### 1.1.3 Doping

Online YouTube lectures:

[102N. Basic Solid-State Physics: Doping, Carrier Density, Distributions](#)

Semiconductor materials are hardly useful in their pure form. The carrier concentrations can be manipulated by introducing dopants in small amounts.

Elements from columns III and V of the periodic table of elements are usually used as dopants.

III	IV	V
B	C	N
Al	Si	P
Ga	Ge	As
In	Sn	Sb

Fifth column elements have five electrons in their outer shell. When an element from column V is introduced into the semiconductor material, four of its outer shell electrons form covalent bonds with 4 adjacent semiconductor atoms, however there will be one extra electron attached to the column five element, as shown in Figure 1.15.

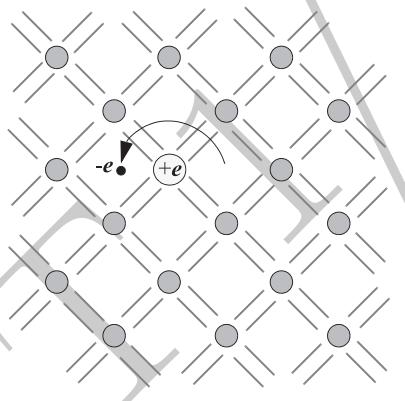


Figure 1.15: An electron loosely attached to an ionized dopant.

To gain an insight to how tightly this fifth electron is bonded to the fifth column atom (now a positive ion) we use the Bohr's model to calculate the binding energy. The combination of the electron and the positive ion looks like a hydrogen atom, but in silicon instead of vacuum. Therefore, the energy levels obtained earlier will be scaled by the square of relative permittivity of the semiconductor. For silicon, we have  $\epsilon_r = 11.7$ , and therefore

$$E_{1,V} \approx \frac{-13.6eV}{\epsilon_r^2} = -0.1eV \quad (1.9)$$

As can be seen, this bonding energy is small compared to the bandgap energy of the silicon ( $E_g = 1.11eV$ ), and hence this fifth electron is bonded much more weakly than the other electrons. In practice at room temperature, this electron is most probably free and in the conduction band. For this reason the fifth column elements are generally referred to as *donors*. They donate electrons to the lattice.) Note that unlike the case of a silicon atom losing an electron, the donor does not leave behind a hole. It rather leaves a localized, immobile positive ion.

Based on the above approximation for the bonding energy of the fifth electron, we have shown the energy band picture for this case in Figure 1.16. From

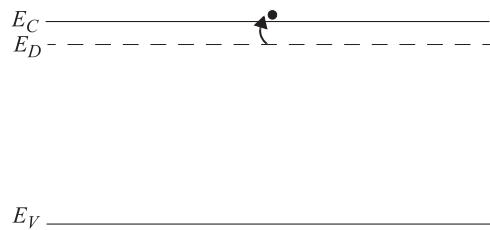


Figure 1.16: Discrete energy levels associated with donor atoms.

now on we will show the valence and conduction energy bands only with lines which designate their borders. Note that the energy band associated with the donor is shown with a broken line to emphasize the localized nature of the donor ions. (In other words, positive donor ions cannot move like holes).

Now if an element from the third column is introduced, it will only have 3 electrons in its outer shell. A free electron can easily fill this opening for the third column atom to form four covalent bonds with adjacent atoms and become a negative ion, as shown in Figure 1.17. This electron is provided by another covalent bond breaking and contributing one electron to the third column element. Note that it is much easier for an electron to break a bond and form another bond than just breaking a bond and becoming a free electron<sup>1</sup>. Due to their tendency to accept free electrons, the third column elements are generally referred to as *acceptors*.

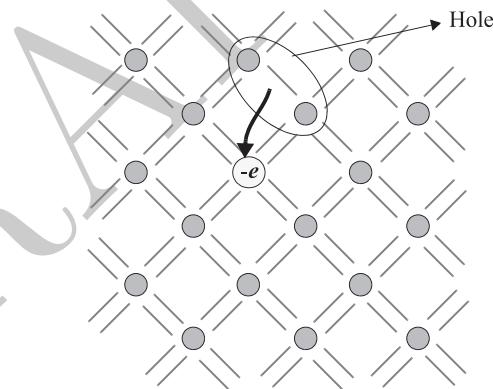


Figure 1.17: An electron captured by an acceptor atom, leaving behind a hole.

In the band diagram picture, acceptors correspond to energy levels close to

---

<sup>1</sup>This is somewhat similar to the quantum mechanical phenomenon of Tunneling.

the valence band. At room temperature most of the acceptor sites are occupied, leaving holes in the valence band. This process is shown in Figure 1.18.

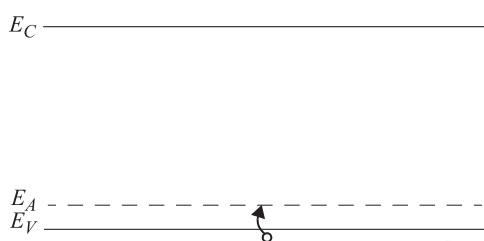


Figure 1.18: Discrete energy levels associated with acceptor atoms.

Again, the energy level associated with the acceptor is shown with a dashed line to emphasize the lack of mobility in the acceptor sites. As the temperature is raised above 0K, more electrons have enough energy to jump from donor sites to the conduction band and from the valence band to the acceptor sites. If we plot the number of free carriers vs. absolute temperature for a pure semiconductor together with the number of carriers in a similar type of semiconductor doped with  $N_D$  donor atoms per unit volume (or  $N_A$  acceptor atoms per unit volume), we observe three different regions shown in Figure 1.19.

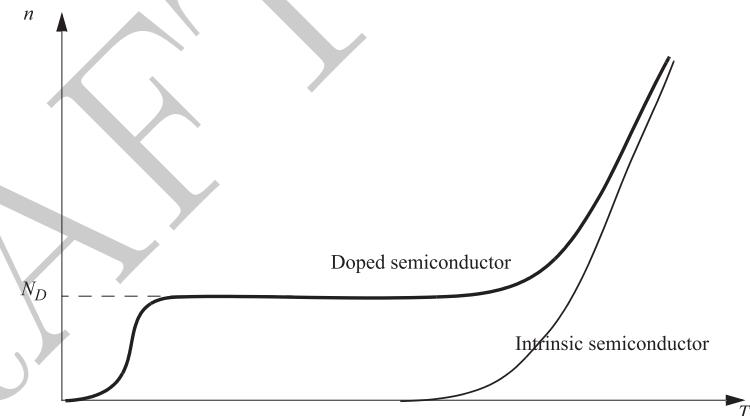


Figure 1.19: Carrier concentration vs. temperature for a doped semiconductor.

At low temperatures the electrons do not have enough thermal energy to jump from the donor site to the conduction band. As temperature is raised, more electrons find enough thermal energy to jump to the conduction band. Therefore, the number of free carriers increases with temperature. Once all of the dopant sites are ionized, the carrier density stays constant. This will continue until the number of dopant induced electrons becomes comparable to

the number of free carriers in the pure semiconductor at that temperature, i.e., the electrons that have enough energy to jump the bandgap. After this point, the thermal energy is large enough to generate more electron-hole pairs than the number of electrons induced by the donors. At high temperatures, the semiconductor simply behaves like an intrinsic (undoped) semiconductor. For electronic devices, it is desirable to operate the semiconductor material in the region with the number of carriers constant with temperature. In this region the number of free carriers is well-controlled and is mainly determined by the number of dopant atoms per unit volume of the semiconductor<sup>2</sup>.

### 1.1.4 Carrier Distribution

Electrons are Fermions and therefore they follow Fermi-Dirac distribution:

$$f(E) = \frac{1}{1 + e^{(E-E_f)/kT}} \quad (1.10)$$

where  $k$  is the Boltzmann constant,  $T$  is the absolute temperature and  $E_f$  is the Fermi energy (or Fermi level). The Fermi function of (1.10) is simply the probability that an electron fills a state with an energy  $E$ . The probability of a state with an energy  $E_f$  being filled with an electron is exactly 1/2. Note that the probability distribution function of (1.10) is only valid under thermal equilibrium conditions. A plot of the Fermi-Dirac distribution is shown in Figure 1.20.

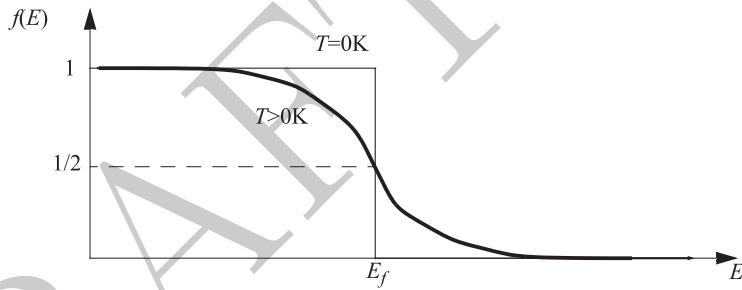


Figure 1.20: Fermi-Dirac distribution.

For  $(E - E_f) > 3kT$  we have  $e^{(E-E_f)/kT} \gg 1$  which implies that Fermi distribution can be approximated using the Boltzmann distribution:

$$f(E) \approx e^{-(E-E_f)/kT} \quad (1.11)$$

Also when  $(E - E_f) < -3kT$ , we have

$$f(E) \approx 1 - e^{(E-E_f)/kT} \quad (1.12)$$

---

<sup>2</sup>It should be evident from this argument that to make semiconductor devices capable of operating at high temperatures, we must incorporate semiconductor materials with higher bandgap voltages. An example of such a semiconductor material is GaN.

We will use this Boltzmann approximation in our discussion of the device physics, which is depicted in Figure 1.21.

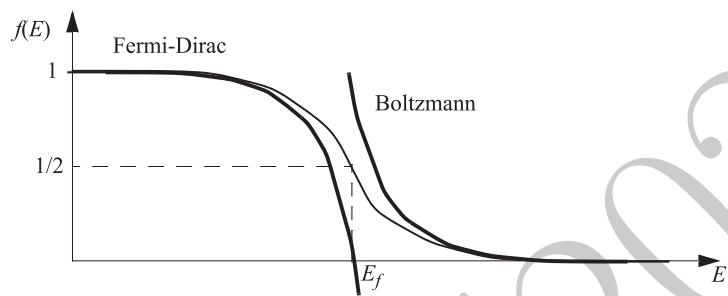


Figure 1.21: Boltzmann distribution approximation to Fermi-Dirac distribution.

One useful result arising from the Boltzmann distribution is that the number of electrons (or holes) with thermal energy greater than an arbitrary energy level,  $E_1$ , is proportional to  $\exp(-E_1/kT)$ , i.e.,

$$f(E > E_1) \propto e^{\frac{E_f - E_1}{kT}} \propto e^{\frac{-E_1}{kT}} \quad (1.13)$$

This is an important relation and we will later use it to derive the current voltage characteristics of junction diodes and bipolar junction transistors.

Based on this expression, the number of electrons in the conduction band will be proportional to  $\exp[(E_f - E_c)/kT]$ . The proportionality constant is shown with,  $N_C$ , and therefore

$$n = N_C \cdot e^{\frac{E_f - E_c}{kT}} \quad (1.14)$$

Similarly, for holes we can write,

$$p = N_V \cdot e^{\frac{E_v - E_f}{kT}} \quad (1.15)$$

where  $N_V$  is another proportionality constant. Parameters,  $N_C$  and  $N_V$ , are called the effective density of states at the edge of conduction and valence band, respectively. For an intrinsic (undoped) piece of semiconductor, the number of thermally generated holes and electrons should be equal and therefore,  $n = p$ . Referring to the Fermi level of the intrinsic semiconductor as  $E_i$ , we can rewrite (1.14) and (1.15) as

$$n_i = N_C e^{\frac{E_f - E_i}{kT}} = N_V e^{\frac{E_v - E_i}{kT}} \quad (1.16)$$

where  $n_i$  is the carrier density in the pure semiconductor and is a function of temperature. Combining (1.14) and (1.15) with (1.16), we obtain the following important relations:

$$\begin{aligned} n &= n_i e^{\frac{E_f - E_i}{kT}} \\ p &= n_i e^{\frac{E_i - E_f}{kT}} \end{aligned} \quad (1.17)$$

This means that if  $E_f$  is above  $E_i$  we have more electrons and fewer holes and vice versa. Equation (1.17) also implies that

$$n \cdot p = n_i^2 \quad (1.18)$$

### 1.1.5 Motion of Free Carriers

Online YouTube lectures:

#### [103N. Carrier Movement in Semiconductors, Drift and Diffusion](#)

As mentioned earlier, the electrons and holes move in the lattice due to their thermal energy. To gain insight into the thermal velocity of charge carriers at room temperature, let us perform a simple calculation. Equipartition theorem of statistical physics suggests that an electron in equilibrium with the lattice at a temperature,  $T$ , has  $kT/2$  energy per degree of freedom. Therefore, we can write

$$\frac{1}{2}m_n^*v_{th}^2 = \frac{3}{2}kT \quad (1.19)$$

where,  $m_n^*$  is the effective mass of the electron and incorporates the effect of all the interactions between the electron and the lattice in a single constant. The effective mass of electrons in silicon is only 0.26 times the mass of a free electron. Using this effective mass in (1.19) we calculate a room mean squared (rms) thermal velocity of approximately  $2 \times 10^5 m/s$ . Intuitively, the reason that this average speed is so high can be traced back to the fact that electrons are very light. Since irrespective of their mass they will have a thermal kinetic energy of  $kT/2$  per degree of freedom, this will translate to a very high velocity for a light particle, such as electron.

Let us assume that the root mean average has the same order of magnitude as the mean<sup>3</sup>. At temperatures above 0K, electrons will collide with the lattice and exchange energy with it. The average time between two collisions with the lattice is called mean scattering time and is shown with<sup>4</sup>  $\tau_c$ .

Now if an electric field is applied to the semiconductor, it exerts force on charge carriers. This is shown in Figure 1.22. According to Newton's second law, the change in the momentum of the electron in the direction of the field is equal to the impulse applied to it. Impulse is defined as the integral of the force over the period of time it is exerted. Since the electric field is assumed to be constant, the change in the momentum will be given by

$$m_n^*\Delta v_1 = -qE\tau_c \quad (1.20)$$

where  $\Delta v_1$  is the change in the component of electron velocity along the electric field. The average *drift velocity*,  $v_d$ , which is the average drift component of the electron velocity is therefore approximately half of  $\Delta v_1$ . The drift velocity for small values of electric field is very small compared to the thermal velocity of

<sup>3</sup>This assumption is not generally true, but good enough for our order of magnitude calculation

<sup>4</sup>In intrinsic silicon at room temperature,  $\tau_c \approx 0.4ps$

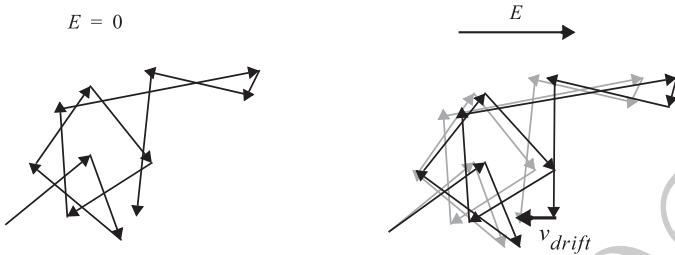


Figure 1.22: Scattered motion of a free charge carrier in the lattice in the absence and presence of an external electric field.

the electron. The drift component can be viewed as a small perturbation on the thermal movements of the electrons, as illustrated in Figure 1.22.

It is noteworthy that although an electron in free space accelerates in the presence of an external electric field, the electrons in the lattice reach a constant average drift velocity,  $v_d$ . This is due to numerous and frequent collisions of the electron with the lattice that randomizes the direction of the gained momentum. In other words, the lattice converts the electric energy to thermal energy, while keeping constant the drift velocity of the charge carriers. As can be seen from (1.20), the drift velocity is proportional to the electric field as long as it is small compared to the thermal energy. Equation (1.20) can be rewritten for electrons as

$$v_d = -\mu_n E \quad (1.21)$$

where

$$\mu_n = \frac{q\tau_c}{2m_n^*} \quad (1.22)$$

is called the mobility of the electron. A similar constant,  $\mu_p$  can be defined for holes. However, (1.21) will have a positive sign for holes. The mobility of holes is usually smaller than the mobility of the electrons. This is due to the larger effective mass of the holes compared to the electrons. Going back to our water filled cylinders, the bubble in the lower cylinder cannot move as easily as the droplet in the upper one, since it experiences a drag force due to the liquid. Therefore, it is not as ‘mobile’ as the droplet.

If electrons have a uniform density,  $n$ , per unit volume and if they move at an average velocity of  $v_d$ . The number of electron passing a surface with an area of unity per unit time is  $nv_d$ , as shown in Figure 1.23. Therefore the electron current density is given by

$$J_n = -nqv_d = nq\mu_n E \quad (1.23)$$

The total current in a semiconductor is carried by both electrons and holes and is thus given by

$$J = q(n\mu_n + p\mu_p)E \quad (1.24)$$

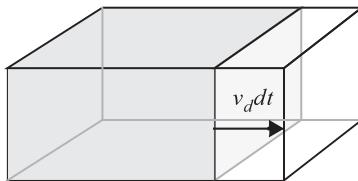


Figure 1.23: Current flow in a slab of semiconductor.

and hence the conductivity of the material is

$$\sigma \equiv \frac{1}{\rho} = q(n\mu_n + p\mu_p) \quad (1.25)$$

The linearity of the drift velocity with the applied electric field does not hold for large values of  $E$ . Remember that one assumption leading to this behavior was that the drift velocity is small compared to the thermal velocity of the electrons. As the electric field becomes larger, electrons gain more momentum between collisions but they due to their speed becoming comparable to the thermal velocity,  $\tau_c$  begins to decrease, reducing the mobility. This prevents electrons from gaining drift velocities larger than the thermal velocity. Therefore, velocity reaches a plateau for large values of  $E$ , as seen in Figure 1.24. This phenomenon is known as velocity saturation and plays an important role in

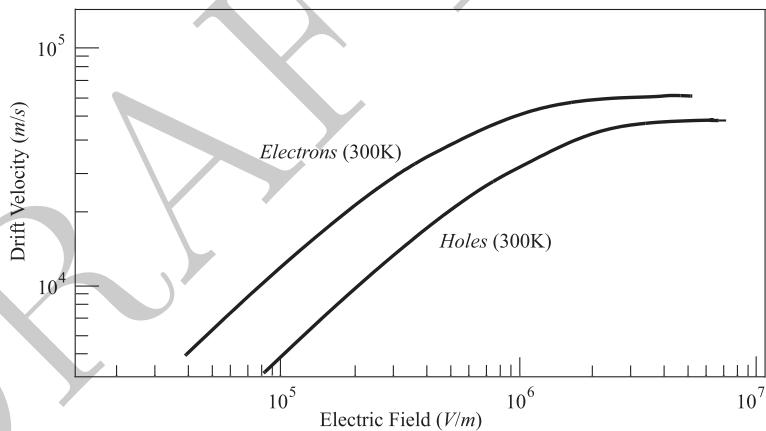


Figure 1.24: Carrier drift velocity vs. electric field showing velocity saturation at high electric field.

short channel metal-oxide-semiconductor (MOS) field effect transistors (FET), as we will see in Section 1.5.

## 1.2 Junction Diode

Online YouTube lectures:

[104N. PN Junction, Depletion Region, Diode Equation](#)

A junction diode consists of two pieces of *n*- and *p*-type semiconductors touching each other. Practically the junction is made out of a single piece of semiconductor material with different levels of doping at different locations. First let us discuss the thermal equilibrium behavior of a *pn* junction.

### 1.2.1 Thermal Equilibrium Behavior

At  $T = 0K$ , neither the acceptors in the *p*-type material, nor the donors in the *n*-type are ionized and the material and band diagrams will look like Figure 1.25. As the temperature is raised above  $0K$ , donor sites release free electrons

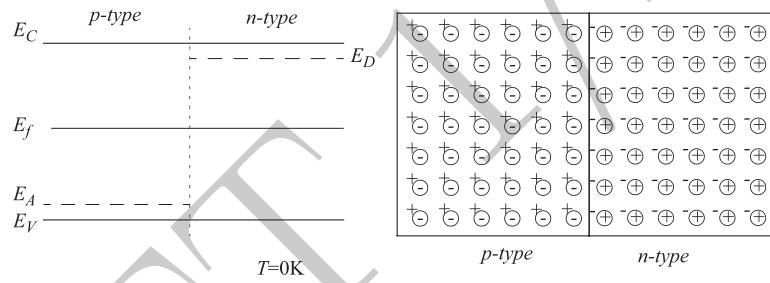


Figure 1.25: The equilibrium energy band diagram of the pn junction at  $0K$ .

on the *n*-side and acceptor sites absorb free electrons (or 'release holes'). These electrons and holes are released due to the thermal energy.

The thermal energy of the crystal lattice also makes these free electrons and holes move around. The thermal movement of each free carrier is totally random. During their random movements, electrons may end up in the *p* region and vice versa for holes. Once on the opposite side, they meet a lot of free carriers of the opposite charge and have a large chance of recombination. Due to this recombination, the area close to the junction becomes depleted of free carriers and hence is called the *depletion region*.

As electrons and holes recombine, they leave positive and negative immobile ions on the edge of the junction in the *n*- and *p*-type material, respectively. The ions on the opposite sides of the junction form an electric field that opposes further movement of the electrons to the *p*-side and holes to the *n*-side, as illustrated in Figure 1.26. The motion due to the thermal energy of electrons and holes reaches a balance with motion in the opposite direction due to the established electric field. As the temperature is raised to room temperature, all the donor and acceptor sites are ionized and the band diagram will look like Figure 1.27.

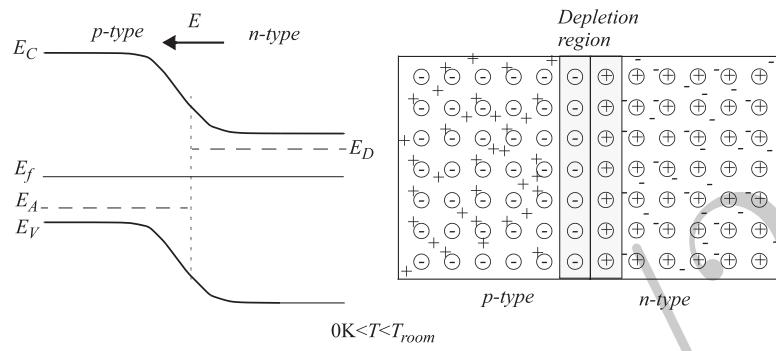


Figure 1.26: The equilibrium energy band diagram of the pn junction at  $0K < T < T_{room}$ .

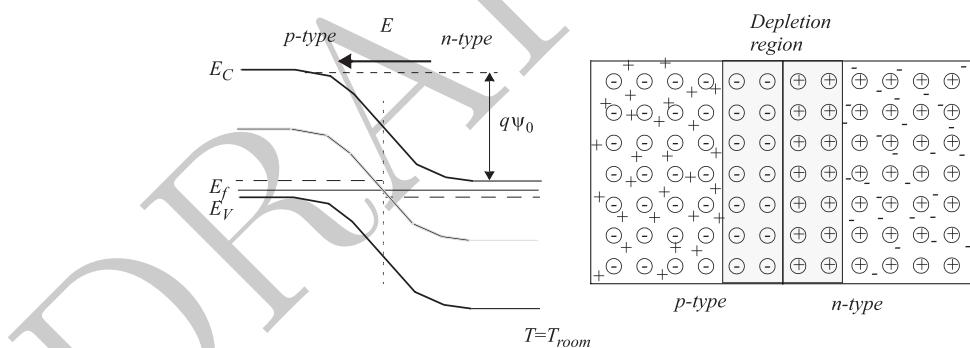


Figure 1.27: The equilibrium energy band diagram of the pn junction at  $T_{room}$ .

The established electric field corresponds to a built-in electric potential  $\psi_0$  which is the sum of the potential drop on the p-side and the n-side, i.e.,  $q\psi_0 = E_{i,n} - E_{i,p}$ , where  $E_{i,n}$  and  $E_{i,p}$  are the energy shift from the junction to the n and p sides. We note that at room temperature all dopants are ionized, and therefore, the density of electrons on the n-side is  $n = N_D$  and similarly the density of the holes on the p-side is  $p = N_A$ . Using (1.17) we can calculate

$$\psi_0 = \frac{E_{i,n} - E_{i,p}}{q} = \frac{kT}{q} \ln \left( \frac{N_A N_D}{n_i^2} \right) \quad (1.26)$$

The electrons in the conduction band of the *n*-type material and the holes in the valence band of the *p*-type material have approximately Boltzmann distributions. Hence in the steady-state the number of electrons that have enough energy to pass the built-in potential barrier is proportional to  $e^{-q\psi_0/kT}$ , where  $q\psi_0$  is the height of the barrier (Figure 1.28). The same is true for holes, but

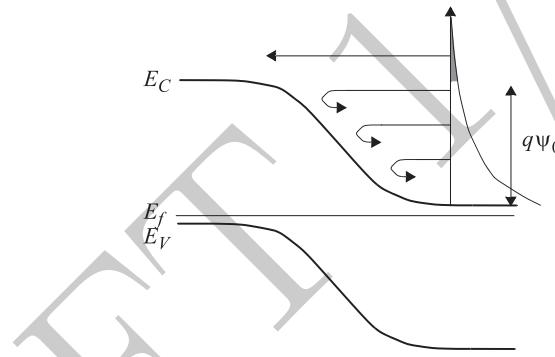


Figure 1.28: Only the electrons with thermal energy higher than the barrier height can cross.

let us concentrate on electrons for the time being.

Although the depletion region has a small number of electrons and holes, every so often an electron-hole pair are created by the thermal generation. When this happens the electric field pushes the hole into the p-region and the electron in the n-region. In the absence of an external potential difference applied to the junction, the current due to the thermal energy of electrons and holes is exactly equal to, but in the opposite direction of, the current caused by the built-in electric field in conjunction with thermal generation. This condition is necessary as no net current flows in the junction.

### 1.2.2 Effect of External Electric Potential

Now assume that an external forward potential is applied to the junction as shown in Figure 1.29. This forces more electrons into the n-region and more holes into the p-region and thereby shrinks the depletion region and lower the

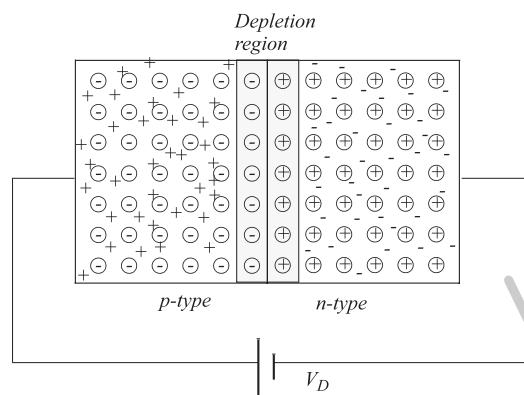


Figure 1.29: pn junction under a forward bias.

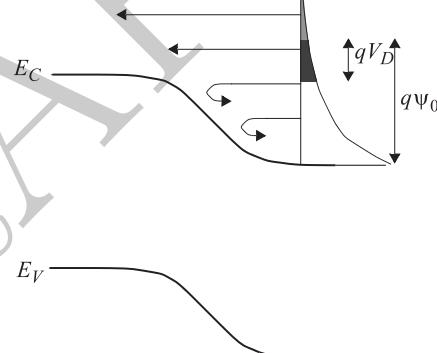


Figure 1.30: Lowered potential barrier due to forward bias allows additional electrons with lower thermal energy to cross.

potential barrier, as shown in Figure 1.30. Now a larger number of electrons have enough thermal energy to pass the barrier. The number of these electrons is proportional to  $e^{-q(\psi_0 - V_D)/kT}$  where  $q(\psi_0 - V_D)$  is the new height of the barrier. Assuming that the current due to the electric field stays the same as when no external field was applied<sup>5</sup>, the net current will be proportional to the number of extra electrons that have enough thermal energy to pass the barrier. These are the electrons that have energies between  $q(\psi_0 - V_D)$  and  $q\psi_0$ , whose number is

$$N_e[q(\psi_0 - V_D) < E < q\psi_0] \propto [e^{-\frac{q(\psi_0 - V_D)}{kT}} - e^{-\frac{q\psi_0}{kT}}] \propto [e^{\frac{qV_D}{kT}} - 1] \quad (1.27)$$

The same proportionality holds for holes. Since the net current is proportional to the sum of the net number of electrons and holes that pass the barrier, it can be written as

$$I(V_D) = I_S(e^{qV_D/kT} - 1) \quad (1.28)$$

where  $I_S$  is a proportionality constant<sup>6</sup>. The coefficient  $kT/q$  is often abbreviated as  $V_T$  and has a value of 25.8mV at 300K (room temperature).

Now we look at the case of a reverse-biased junction. Applying an electric potential in the opposite direction will increase the width of the depletion region, as depicted in Figure 1.31. This increase in the width of the depletion region is

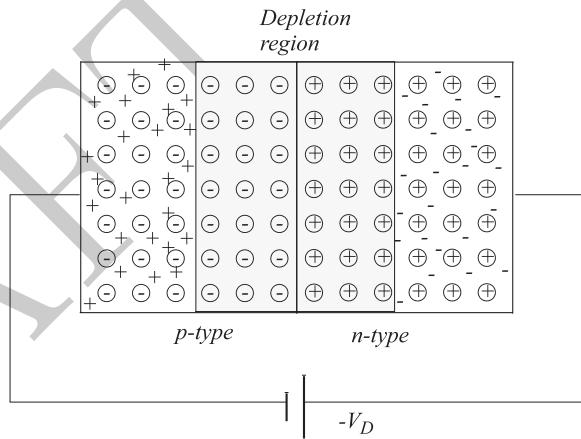


Figure 1.31: pn junction under reverse bias.

necessary to accommodate the larger potential difference across the junction by

<sup>5</sup>This assumption is not strictly valid, since the electric field at the junction is changed; however, it does not have a significant effect on the final result.

<sup>6</sup>As one can easily see from the argument,  $I_S$  itself must be proportional to the area of the junction. In other words, doubling the size of the junction would result in doubling of the current for the same voltage. An easy way to see this is by noting that a junction with double the area can be made by placing two of the original junction side-by-side. This new compound junction, however, will carry twice the current.

raising the built-in electric field. This will be equivalent to an increase in the height of the potential barrier (Figure 1.32), which in turn results in a reduction

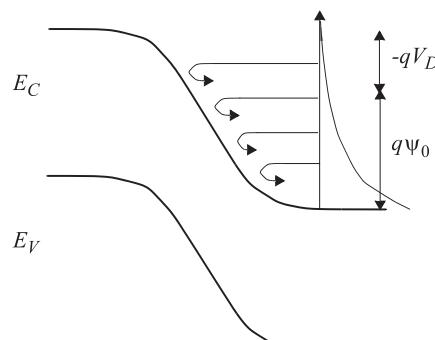


Figure 1.32: Higher potential barrier due to reverse bias allows fewer electrons to cross the barrier.

of the number of electrons having enough energy to pass the potential barrier.

Again assuming that the current due to the electric field does not change significantly, the number of electrons having enough thermal energy to pass the barrier becomes negligible compared to the current due to the electric field. Therefore, the net reverse current will be  $-I_S$ . Equation (1.28) also predicts this behavior since in its derivation no assumption about the sign of  $V_D$  was made. Thus, (1.28) is valid for both forward- and reverse-biased diode.

### 1.2.3 Junction Capacitance

Online YouTube lectures:

[\*\*105N. PN Junction, Junction Capacitance, Doping Profile\*\*](#)

The depletion region of a *pn* junction resembles a capacitor in the sense that negative and positive charges are stored on the opposite sides of the junction<sup>7</sup>. The charge depends on the potential drop across the junction as should be the case in a capacitor. To calculate the capacitance attributed to the junction, we need to calculate the width of depletion region as a function of  $V_D$ . The capacitance per unit area of a *pn* junction is given by

$$C = \frac{\epsilon_S}{x_n + x_p} \quad (1.29)$$

where  $\epsilon_S$  is the permittivity of the semiconductor material and  $x_n$  and  $x_p$  are the widths of the depletion regions in the *n*- and *p*-type material, as shown in

---

<sup>7</sup>Although these charges are not mobile, a change in the voltage across the diode results in a change in the total amount of charge in this region. This change is caused by the flow of free electrons and holes into (or out of) the depletion region. This flow is similar to the charge flow in a parallel plate capacitor. It should be clear from this argument that the definition of the junction capacitance is a small-signal one.

(Figure 1.33). We continue with the assumption that the doping profile changes

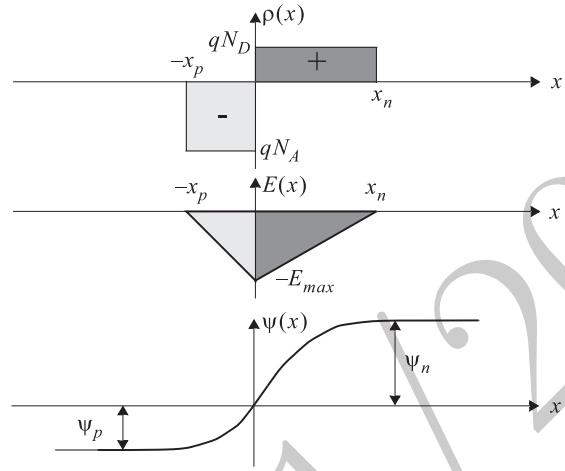


Figure 1.33: Electric charge density, electric field, and electric potential profile across an *abrupt* pn junction.

very rapidly at the junction. To maintain charge neutrality, the total charge of the ions in the *n*-type material should be equal to the charge of the ions in the *p*-type material, i.e.,

$$x_n N_D = x_p N_A \quad (1.30)$$

where  $N_A$  and  $N_D$  are the doping levels in the *p*-type and *n*-type material, respectively. The electric field can be calculated at any point using Gauss's law. The electric field is maximum at the junction and its maximum value is

$$E_{max} = \frac{q N_A x_p}{\epsilon_S} = \frac{q N_D x_n}{\epsilon_S} \quad (1.31)$$

This expression for the electric field can be rewritten more symmetrically as

$$E_{max} = \frac{q}{\epsilon_S} \cdot \frac{N_A N_D}{N_A + N_D} (x_n + x_p) \quad (1.32)$$

The electric potential is negative the integral of the electric field across the junction. In other words, the electric potential drop across the junction is equal to the area under the electric field plot. Thus, the potential drop on the *n*- and *p*-side of the depletion region are

$$\begin{aligned} \psi_n &= \frac{E_{max} x_n}{2} \\ \psi_p &= \frac{E_{max} x_p}{2} \end{aligned} \quad (1.33)$$

Therefore the potential drop across the junction is given by the sum of these two terms, i.e.,

$$\psi_0 - V_D = \psi_n + \psi_p = \frac{E_{max}(x_n + x_p)}{2} = \frac{q}{2\epsilon_S} \cdot \frac{N_A N_D}{N_A + N_D} (x_n + x_p)^2 \quad (1.34)$$

Solving for the depletion region width, we have:

$$x_d \equiv x_n + x_p = \left[ \frac{2\epsilon_S}{q} \left( \frac{1}{N_A} + \frac{1}{N_D} \right) (\psi_0 - V_D) \right]^{\frac{1}{2}} \quad (1.35)$$

A few things can be learned from (1.35). First, the width of the depletion region, and hence the junction capacitance, is dominated by the side with lighter doping (smaller  $N$ ). Second, the width of the depletion region grows as the square root of the potential drop across the junction for an abrupt junction. This should be intuitive at this point, as in the graph of the electric field vs. position, doubling of the junction width will quadruple the area under the graph (the slope of the curve is controlled by the doping levels and is therefore constant). Using (1.29) and (1.35), the junction capacitance of an abrupt junction can be easily calculated to be:

$$C_j = \left[ \frac{q\epsilon_S}{2 \left( \frac{1}{N_A} + \frac{1}{N_D} \right) (\psi_0 - V_D)} \right]^{\frac{1}{2}} = \frac{C_{j0}}{(1 - \frac{V_D}{\psi_0})^{\frac{1}{2}}} \quad (1.36)$$

where  $C_{j0}$  is the junction capacitance for zero external bias. As can be seen, the junction capacitance is voltage dependent and therefore behaves as a nonlinear capacitor. Note that the above 1/2 exponent of the denominator directly arises from the abrupt junction assumption. We can use Figure 1.34 to see how the exponent will be affected by the doping profile let us consider the case of a linearly graded junction. In a linearly graded junction the doping profile linearly changes with  $x$ . Since the charge density changes linearly, the electric field, which is the integral of charge density, will have a quadratic dependence on  $x$  and hence the electric potential will be proportional to  $x^3$ . In this case, the width of the depletion region will be proportional to  $(\psi_0 - V_D)^{\frac{1}{3}}$  and therefore the capacitance will be given by

$$C_j = \frac{C_{j0}}{(1 - \frac{V_D}{\psi_0})^{\frac{1}{3}}} \quad (1.37)$$

In practice, different junctions have different doping profile, hence the exponent of the denominator can have a value different from the above values. This exponent is usually shown as  $m$ .

### 1.3 Bipolar Junction Transistor

Online YouTube lectures:

[106N. Bipolar Junction Transistor, basic operation, current flow properties, doping Profile](#)

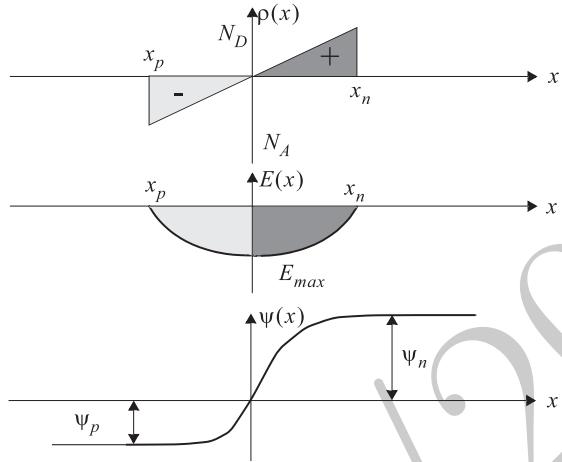


Figure 1.34: Electric charge density, electric field, and electric potential profile across an *linearly graded* pn junction.

To build a useful amplifier or switch, we need more than two terminals. In general, we need a device whose output voltage or current is controlled by the voltage or current of a separate terminal. This isolation of input and output ports is very important in building an amplifier or a logic gate. Thus, the minimum number of terminals is three. We will modify a junction diode so that its current is controlled by the voltage of a different node.

As we saw earlier, in a junction diode carriers are injected from one side of the junction to the other side. The number of carriers having enough thermal energy to pass the barrier is controlled by Boltzmann distribution and depends exponentially on the electrical potential drop across the junction. Once these carriers reach the opposite side, they become *minority carriers*. Electrons are minority carriers in *p*-type material, so are holes in *n*-type material. In a junction diode, these minority carriers gradually recombine with the majority carriers so that after a reasonable distance from the junction, most of the current is carried by the majority carriers. This is shown pictorially in Figure 1.35.

One way to direct the minority carrier current to a third terminal is by absorbing them into a region where these carriers are majority carriers. If the minority carriers are absorbed a short distance away from the junction, only a small number of them are lost due to recombination. This can be achieved by making a sandwich of two semiconductor of the same type and a thin layer of opposite type of semiconductor. The carriers will be injected ('emitted') from the first piece of semiconductor into the small piece in the middle where they are minority carriers. However, once they are 'collected' by the third piece of semiconductor, they are majority carriers again. The collection of these minority carrier can be facilitated by reverse biasing the second junction formed between the second and third pieces of semiconductor. If we make the structure out

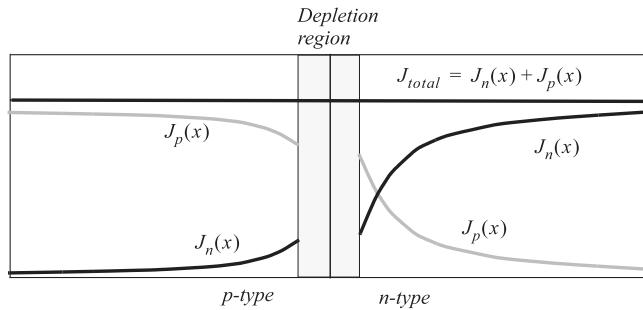


Figure 1.35: Electron, hole, and total current profiles across a forward biased pn junction.

of an *p*-type layer sandwiched between two *n*-type layers, we form a so-called, *npn* transistor. This way, once the electrons make it to the edge of the second depletion region, the built-in electric field of the junction will absorb them into the *n*-type region. The device looks more or less like Figure 1.36

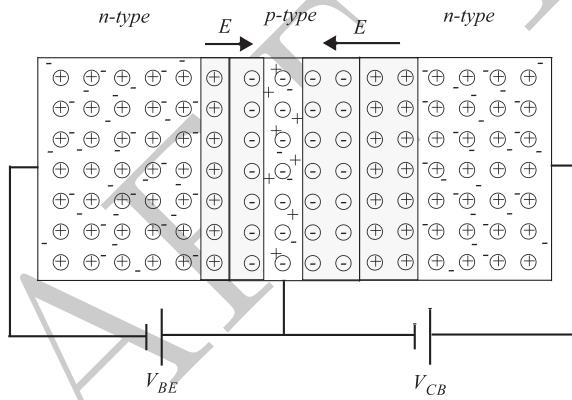


Figure 1.36: An npn transistor with the *base-emitter* junction forward biased and *base-collector* junction reverse biased (Forward Active Region).

The semiconductor region on the left is called *emitter* since it emits charged carriers in the device. The thin region of semiconductor in the middle is called *base*. The semiconductor region on the right is called *collector* as it collects the minority carriers injected into the base. A complementary device, called a *pnp* transistors can also be made out of two pieces of *p*-type semiconductor sandwiching a thin *n*-type slab. These two devices are usually shown with the symbols shown in Figure 1.37.

Now let us look at the band diagram picture of an *npn* transistor, shown in Figure 1.38 We have two *pn* junctions sharing a thin *p*-type base. The emitter

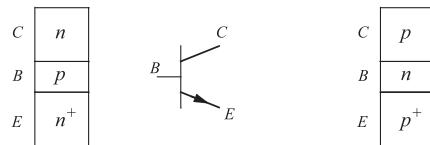


Figure 1.37: The *npn* and *pnp* transistors with their associated symbols.

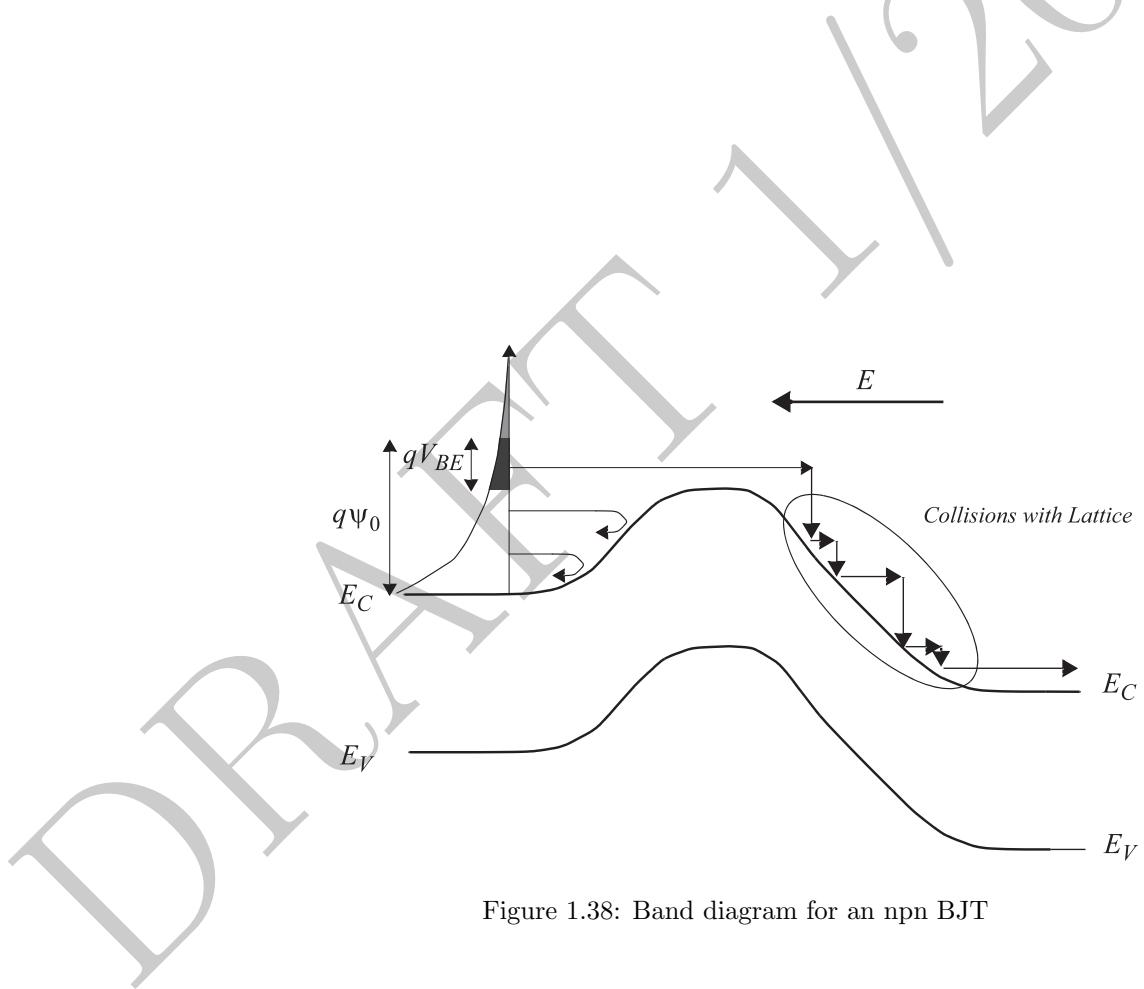


Figure 1.38: Band diagram for an npn BJT

base junction is forward biased so electrons with thermal energies larger than the barrier height can pass the barriers. The number of these electrons, as we discussed earlier, is proportional to  $e^{q(\psi_0 - V_{BE})/kT}$ . Therefore, the electron current carried by these thermal electrons is given by (1.28), replacing  $V_D$  with  $V_{BE}$ .

The base is a hostile environment to the electrons as they are minority carriers there. However, if the base width is kept narrow, most of these electrons make it to the edge of collector-base depletion region where they are pulled into the collector with the aid of the established electric field in this junction. Once in the collector neutral region, the electrons are majority carriers again.

The primary reason for the electrons to go across the base is their thermal energy. As we saw earlier electrons' thermal velocity is very large due to their lightness<sup>8</sup>. Eventually electrons in the collector lose their excess kinetic energy through lattice scattering. If we assume that most of this current is carried by electrons<sup>9</sup>, and that the number of electrons recombining with hole in the base is negligible, the collector current will be given by the same expression as the electron current through the base-emitter junction, i.e.,

$$I_c = I_S \cdot e^{V_{BE}/V_T} \quad (1.38)$$

Note that although this current is controlled by the base-emitter voltage, it is carried mainly by the collector. A typical plot of this transfer characteristic looks as illustrated in Figure 1.39.

We ignored an important effect in the foregoing argument. If the doping levels of the emitter and the base are comparable, only half of the current is carried by electrons and the other half will be carried by holes. Holes will be injected from the base into the emitter, but since we have no mechanism for collecting them, they will recombine in the emitter and will not contribute to the collector current. In this case, the collector current will be about half of the emitter current which will severely limit the performance of the transistor.

People have found a clever, yet simple, solution to this problem. If the emitter is doped much more heavily than the base, the concentration of electrons in the emitter will be much larger than the concentration of holes in the base and there are simply more electrons available than holes. Thus, a large fraction of the current will be carried by the electrons<sup>10</sup>. In fact, we can define an efficiency for this process which we call *emitter injection efficiency* for apparent reasons.

---

<sup>8</sup>In addition to the first order effect of the thermal energy, doping profile in the base can be designed to form an electric field facilitating the movement of the electrons toward the collector. However, this effect is still relatively small compared to the thermal movement.

<sup>9</sup>We will shortly see what conditions are necessary for this assumption to hold.

<sup>10</sup>There is another way to improve the emitter injection efficiency and that is by introducing a different kind of semiconductor in the base with a smaller bandgap. Done properly, this can increase the barrier height for the holes without a one-to-one increase in the barrier height for the electrons. Even a small difference in the barrier height seen by the electrons and holes can result in a considerable difference in the number of electrons and holes injected into the opposite region due to the exponential nature of the thermal energy distribution. This is the basis of the so-called hetero-junction bipolar transistors (HBT).

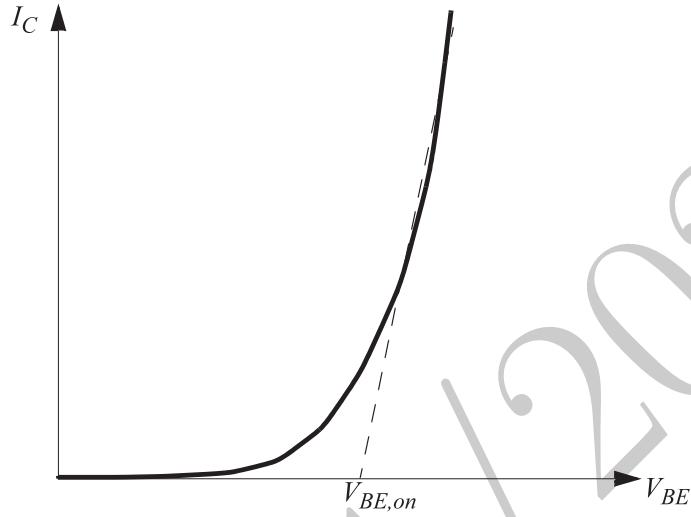


Figure 1.39: A typical  $I_C$  vs.  $V_{BE}$  plot for a BJT

It is the ratio of the emitter electron current to the total emitter current, i.e.,

$$\gamma \equiv \frac{I_{nE}}{I_{nE} + I_{pE}} \quad (1.39)$$

where  $I_{nE}$  and  $I_{pE}$  are the emitter electron and hole currents, respectively. In a well designed integrated circuit transistor,  $\gamma$  is close to 1. Another non-ideality is the recombination of electrons with holes in the base. A fraction of electrons that start at the emitter edge of the base will recombine with the holes in the base before they make it to the collector. The efficiency of this transport process is called *base transport factor*,  $\alpha_T$ , and is defined as the ratio of the number of electrons that make it to the collector to the number of electrons that leave the emitter, i.e.,

$$\alpha_T \equiv \frac{I_{nC}}{I_{nE}} \quad (1.40)$$

where  $I_{nC}$  and  $I_{nE}$  are the electron current at the edge of the collector and emitter, respectively. For a transistor with a thin base,  $\alpha_T$  is close to one. Although both  $\gamma$  and  $\alpha_T$  are close to one in a well-design transistor, a finite number of holes are still injected back into the emitter and also some of the base holes annihilate while recombining with the injected electrons. In steady-state, these lost holes should be replaced by the new holes which are provided by the base current. The overall efficiency of the electron transport phenomenon is given by the product of these two efficiencies, i.e.,

$$\alpha_0 \equiv \frac{I_C}{I_E} = \alpha_T \cdot \gamma \quad (1.41)$$

Due to conservation of electric charge, the net current going into the device should be zero, therefore,

$$I_C + I_B = I_E \quad (1.42)$$

Combining (1.41) and (1.42), we obtain

$$I_C = \frac{\alpha_0}{1 - \alpha_0} I_B = \beta_0 I_B \quad (1.43)$$

where

$$\beta_0 \equiv \frac{\alpha_0}{1 - \alpha_0} \quad (1.44)$$

is an important parameter for a bipolar transistor.

The different components of the dc current in an npn transistor are shown in Figure 1.40.

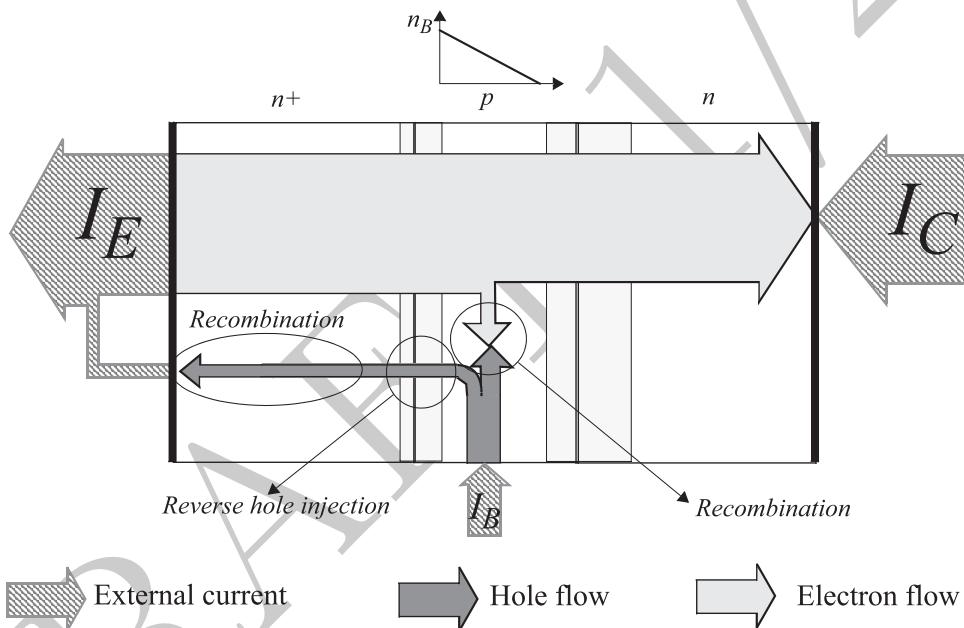


Figure 1.40: Different current components in an *npn* BJT

While the base-collector junction is reversed biased, the transistor can be modeled with an equivalent circuit similar to Figure 1.41. Noting that

$$I_C = \alpha_0 I_E = \alpha_0 I_{ES} \left( e^{V_{BE}/V_T} - 1 \right) \quad (1.45)$$

and comparing it to (1.38) we have

$$\alpha_{dc} I_{ES} = I_S \quad (1.46)$$

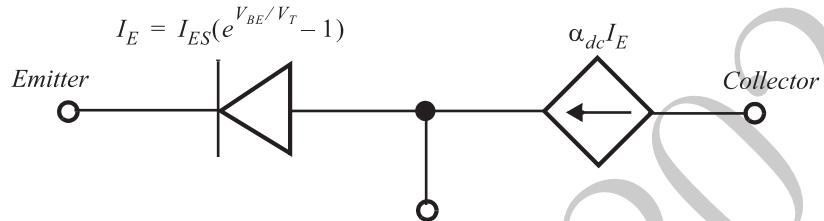


Figure 1.41: Nonlinear model for an *n*p*n* BJT in the forward active region (*base-emitter* forward biased and *base-collector* reverse biased)

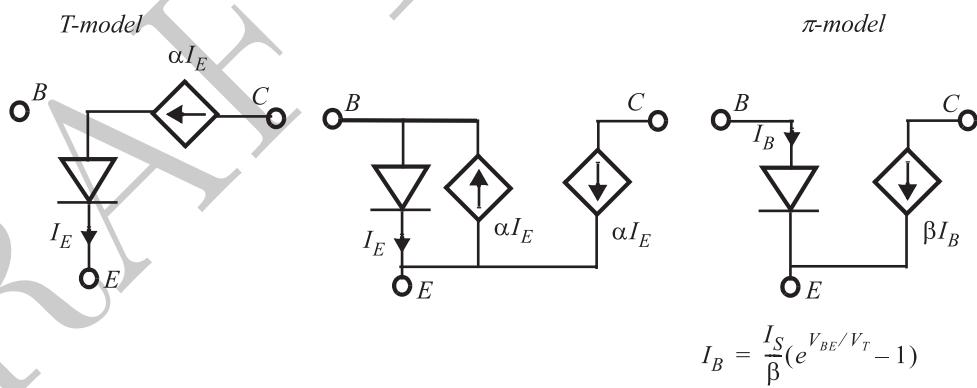


Figure 1.42: Conversion of the nonlinear (large-signal) T-model to the nonlinear (large-signal)  $\pi$ -model

This model known as the large-signal *T-model* can be converted to an equivalent model called *π-model*, through the steps of Figure 1.42. The left most figure shows the original model. The current source can be broken into two equal sources: one going out of the collector and into the base, and the second out of the base into the emitter. Note that the net current into any of the nodes is not changed. Now, the total current in the left branch of the circuit is  $(1 - \alpha_0)I_E = I_B$ , therefore the current through the diode between the base and the emitter will be

$$I_B = \frac{I_S}{\beta} (e^{V_{BE}/V_T} - 1) = \frac{I_C}{\beta} \quad (1.47)$$

Depending on the circuit topology, the *T*- or *π*-model may be more useful.

Online YouTube lectures:

#### [107N. Bipolar transistor: Early effect, Ebers-Moll model, large-signal T- pi-models, dynamics](#)

According to the foregoing analysis, the collector current is independent of the reverse bias across the base-collector junction. The reverse bias was there merely to facilitate the collection of electrons into the collector. However, the collector current of a bipolar junction transistor increases with increasing the collector voltage, as illustrated in Figure 1.43. This effect is known as Early

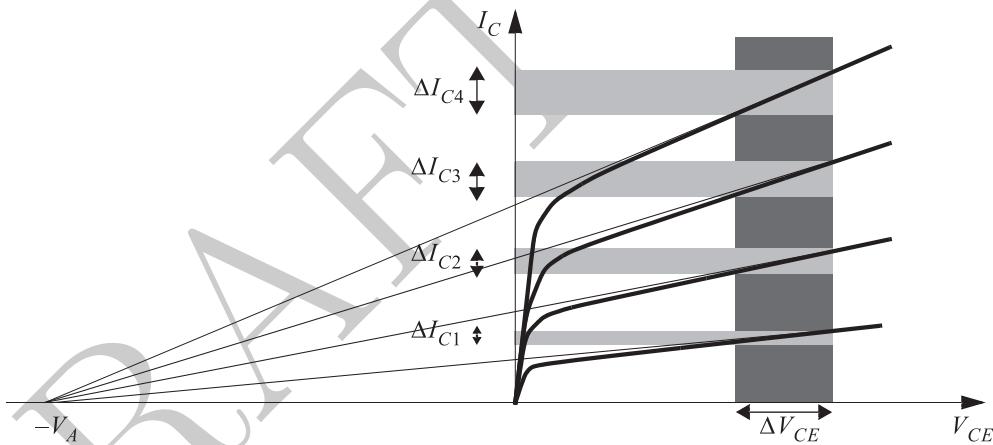


Figure 1.43:  $I_C$  vs.  $V_{CE}$  plots for an npn BJT, demonstrating based width modulation (Early effect)

effect<sup>11</sup> and was first described by James Early in 1952. A simple explanation of this effect follows. The base width affects the number of electrons that can make it to the collector due to their thermal energy. The shorter the base width, the higher the chance of an electron reaching collectors due to its random thermal

<sup>11</sup>It is also referred to as *base width modulation*

movements<sup>12</sup>. Thus, changing the width of the base will change the collector current, increasing  $I_C$  when base width is reduced. We saw earlier that the width of the depletion region of a reversed biased junction increases when the reverse bias is increased. As suggested by Figure 1.44, the effective base-width

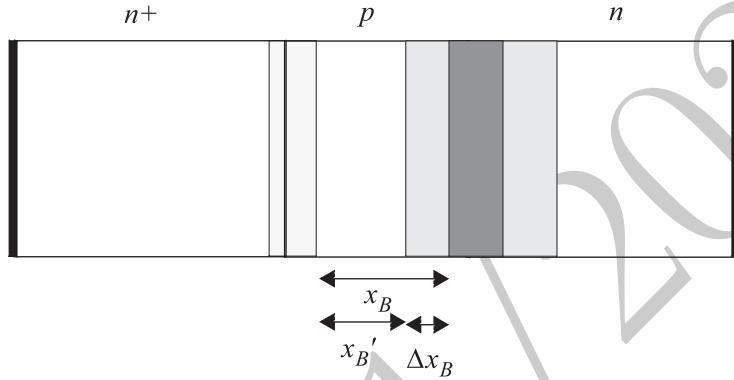


Figure 1.44: Base width modulation due to a change in the *collector-base* reverse bias.

shrinks as the collector-base depletion region extends further into the base due to the larger reverse bias<sup>13</sup>.

As can be seen from the  $I_C - V_{CE}$  graph shown earlier, for larger values of  $I_C$ , the change in  $I_C$  for a given change in  $V_{CE}$  is proportionally larger. This can be easily explained noting that the collector current is controlled by the number of minority carriers (free electrons) in the base. A given change in the effective base width due to the collector voltage change results in the same fractional change in the total base minority charge. Therefore, when  $I_C$  is larger,  $\Delta I_C$  is larger by the same amount. In other words, the ratio of the  $I_C$  to the slope of the  $I_C$  with respect to  $V_{CE}$  is constant to the first order and is called Early voltage<sup>14</sup>:

$$V_A = \frac{I_C}{\frac{\partial I_C}{\partial V_{CE}}} \quad (1.48)$$

Thus, all the tangents to the  $I_C - V_{CE}$  graphs intercept at a point on the  $x$ -axis with a value of  $-V_A$ . Based on this observation we can modify the expression

<sup>12</sup>There is another effect that facilitates the movement of electrons to the collector under a larger base-collector reverse bias. In practice a small part of the electric field extends beyond the depletion region, so applying a larger reverse bias increases the favorable field within the base. Although this effect can be important, the random thermal movements of charge carriers is the primary reason for the operation of the transistor.

<sup>13</sup>There is a second reason for this dependence of  $I_C$  on the  $V_{CE}$  and that is the extension of the electric field of the base-collector depletion region into the base, which facilitates the flow of the electrons.

<sup>14</sup>The reciprocal of this quantity can be thought of as the ratio of changes in collector current with respect to collector-emitter voltage normalized to the absolute value of collector voltage variations.

for the collector current of a bipolar junction transistor to include a term taking Early effect into account, i.e.,

$$I_C = I_S \cdot e^{V_{BE}/V_T} \cdot \left( 1 + \frac{V_{CE}}{V_A} \right). \quad (1.49)$$

As will become more clear later on, we would like to minimize the Early effect in a BJT (bipolar junction transistor). To minimize Early effect we need to lower the ratio of the change in the base width,  $\Delta x_B$ , to the effective base width,  $x_B$ . The first thing that comes to mind is to increase the base width. However, this has a destructive effect on the beta and the high frequency response of the transistor. A better way of increasing the Early voltage is by controlling the doping levels. Noting that the depletion region width is dominated by its width on the side with lighter doping, a larger Early voltage can be obtained by doping the collector side much lighter than the base. Remembering that the emitter had to be doped more heavily to have high emitter injection efficiency, the doping levels should drop from the emitter to the collector in a properly designed bipolar transistor. However this can increase the collector resistance which can result in unnecessary gain and power loss. To avoid this the lightly doped region of the collector is often followed by a highly doped region of the same kind (e.g.,  $n+$  in the case of an npn transistors) to minimize the collector series resistance<sup>15</sup>.

Up to this point, we assumed that the base-collector junction is reversed biased. However, the base-collector junction can become forward biased too. In fact, the role of the emitter and the collector can be reversed by forward biasing the base-collector junction and reverse biasing the emitter-base junction. However, this ‘new’ transistor has several undesirable characteristics. First of all, its emitter has a *lower* doping concentration than its base, therefore, its emitter injection efficiency is very low. Second, a large part of the base-collector depletion region is formed in the base, resulting in severe channel length modulation and a small Early voltage. Thus, both its  $\beta$  and its  $V_A$  are lower. This regime of operation is referred to as *reverse active* and is seldom used.

The transistor’s large signal model can therefore be augmented with the diode and the dependent current source associated with the reverse transistor. This more complete model looks like Figure 1.45. This model is known as Ebers-Moll model of bipolar junction transistor. There are two junctions in the bipolar transistor. Therefore, there are four possible combinations of the status of the junctions, resulting in four different regimes of operation:

---

<sup>15</sup>An HBT discussed in a footnote a few page earlier also helps with the collector resistance, by allowing the base (and hence the collector) to be designed at a higher doping concentrations, without a loss in the emitter injection efficiency. This increase in doping levels lowers the base and the collector series resistance.

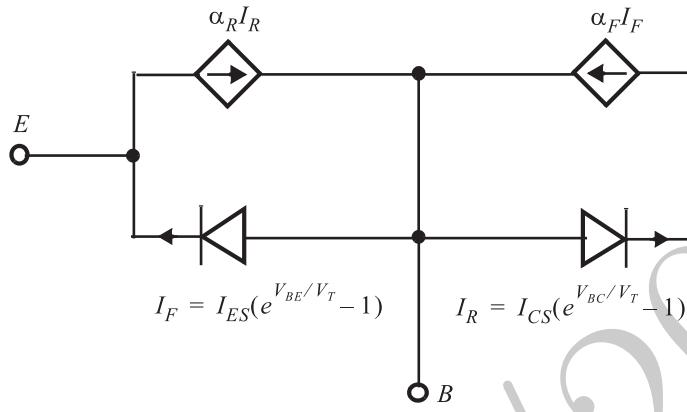


Figure 1.45: Ebers-Moll model for a BJT.

BE BC	Reverse Biased	Forward Biased
Reverse Biased	Cut Off	Forward Active
Forward Biased	Reverse Active	Saturation

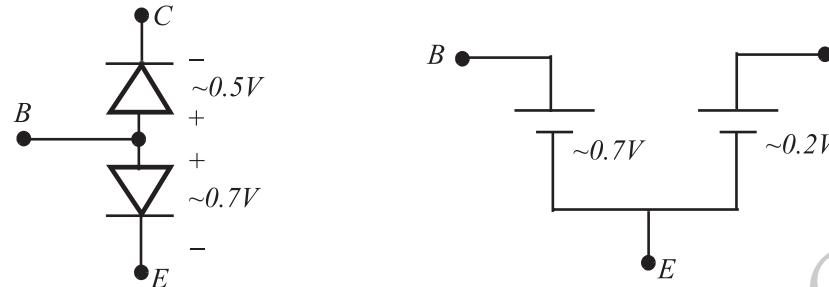
In the cut-off region, both junctions are reversed biased and no significant dc current goes through any of the terminals. In this case, only the junction capacitors will be seen by the external circuitry.

When both junctions are forward biased, the transistor is in saturation. For many purposes, the transistor can be modeled as two back-to-back forward biased diodes in saturation. For silicon transistor the base-emitter junction has a voltage in the range of 0.6V-0.8V with typical values of current. Also the base-collector diode needs to be forward biased by at least 0.4V-0.6V for it carry a significant current. Therefore, in saturation the transistor can be modeled as two constant voltage sources between base-emitter and base-collector, as depicted in Figure 1.46.

### 1.3.1 Transistor's Dynamic Behavior

Up to this point we only investigated the low frequency behavior of the transistors. The high frequency behavior of the bipolar junction transistors is affected by both the junction capacitances and the minority carrier charge in the transistor. We introduced the nonlinear depletion capacitors in a junction before. So let us focus on the effect of minority charge storage in the transistor.

In an *npn* transistor, the collector current is controlled by the amount of electronic charge in the base because the larger the number of electrons in the base, the higher their chance to enter collector-base depletion region due to their random thermal movements. Defining *base transit time*,  $\tau_F$ , as the average time it takes an electron in the base to reach the edge of the collector-base junction



*At the verge of saturation*

Figure 1.46: Simplifier model for BJT in saturation region.

due to thermal movements, the collector current can be expressed as

$$I_C = \frac{Q_F}{\tau_F} \quad (1.50)$$

where  $Q_F$  is the total minority carrier charge in the base<sup>16</sup>. The number of the electrons in the base is affected by two main parameters. First, the number of electrons making it to the collector is smaller than the number of electrons leaving emitter. The remaining electrons have to end up in the base. Therefore, the base minority carrier charge grows with a rate of  $I_E - I_C$ , which is equal to  $I_B$  due to KCL. The second effect is the recombination of electrons and holes, This recombination constantly reduces the base charge,  $Q_F$ . It can be modeled as a relaxation process with a time constant of  $\tau_B$ , since the number of electrons recombining with holes is proportional to the number of available electrons. Therefore, the total base charge is controlled by the following equation:

$$\frac{dQ_F}{dt} = I_E - I_C - \frac{Q_F}{\tau_B} = I_B - \frac{Q_F}{\tau_B} \quad (1.51)$$

where the first term on the right hand side represents the electrons leaving the emitter and not making it to the collector, and the second term represents the electron recombining with the holes in the base. Note that (1.51) does *not* include the effect of the junction capacitors. In steady-state, the changes in the charge is zero and therefore,

$$I_B = \frac{Q_F}{\tau_B} \quad (1.52)$$

<sup>16</sup>In reality,  $Q_F$  consists of two components, the minority carrier charge in the base (electrons in an *npn* transistor) and the minority carrier charge in the emitter (holes in an *npn* transistor). Equation (1.50) is still valid in this more general case, with a different value for  $\tau_F$ . Note that in this case  $\tau_F$  is not exactly the average base transit time for electrons in the base. Nonetheless, the deviation from this value is not large if the emitter injection efficiency is close to 1, i.e.,  $\tau_F$  will be close to the actual average base transit time for electrons if electrons form the dominant part of the  $Q_F$ .

Comparing (1.50) and (1.52), we notice that

$$\beta_0 = \frac{I_C}{I_B} = \frac{\tau_B}{\tau_F} \quad (1.53)$$

Hence knowing  $\beta_0$  and  $\tau_F$  one can predict the dynamic behavior of the transistor to the first order.

### 1.3.2 Small-Signal Model for Bipolar Junction Transistors

In many analog applications, it is desirable to linearize the nonlinear current-voltage transfer characteristic of the bipolar junction transistors to gain design insights. Since the bipolar junction transistor is a voltage-controlled current source, it is most useful to define its transconductance as the derivative of the collector current with respect to the base-emitter voltage. Using (1.38) we obtain

$$g_m \equiv \frac{\partial I_C}{\partial V_{BE}} = \frac{I_C}{V_T} = \frac{1}{r_m} \quad (1.54)$$

where  $I_C$  is the dc collector current at the operation point and  $V_T = kT/q$ . This transconductance can be interpreted as the slope of the  $I_C$ - $V_{BE}$  curves (Figure 1.47) of the transistor at the operating current of interest.

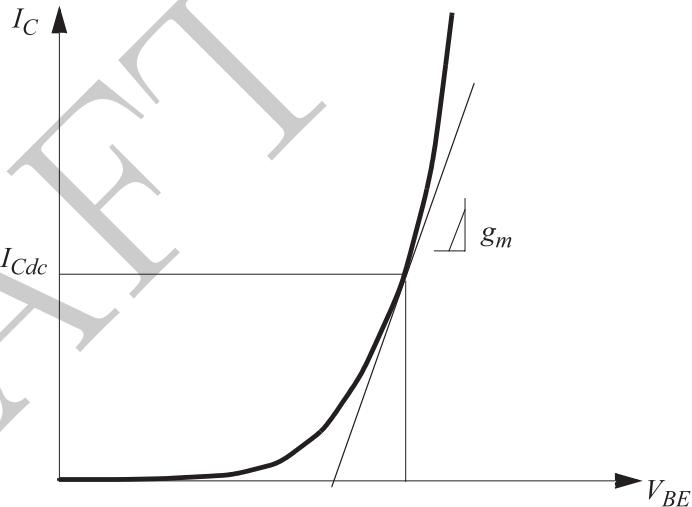


Figure 1.47: The local slope of the  $I_C$  vs.  $V_{BE}$  around the point of operation provides the transconductance  $g_m$ .

The small-signal input resistance at the base terminal of the transistor can be expressed as

$$r_\pi \equiv \left( \frac{\partial I_B}{\partial V_{BE}} \right)^{-1} = \beta_0 \left( \frac{\partial I_C}{\partial V_{BE}} \right)^{-1} = \frac{\beta_0}{g_m} = \beta r_m \quad (1.55)$$

The small signal resistance between the collector and emitter is given by the derivative of the collector current with respect to the collector-emitter voltage. Equation (1.48) results in

$$r_o = \left( \frac{\partial I_C}{\partial V_{CE}} \right)^{-1} \approx \frac{V_A}{I_C} \quad (1.56)$$

The small signal transistor model up to this point is illustrated in (Figure 1.48). These three elements are usually the most dominant low frequency components

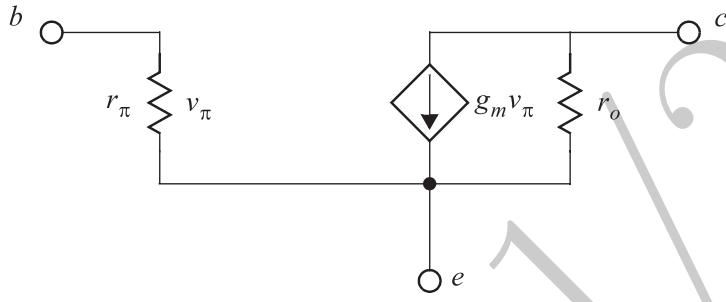


Figure 1.48: The basic hybrid- $\pi$  model for a BJT.

of the above small signal model also known as *hybrid- $\pi$*  model. There are other elements in the hybrid- $\pi$  model that we will calculate here.

Note that although the hybrid- $\pi$  model was derived for an *npn* transistor, it will be the same for a *pnp* transistor as well, since it only deals with the incremental variations of the voltages and currents and not the DC values.

A change in the collector-emitter voltage will result in a change in the collector current which corresponds to a change in the base current. This effect can be characterized by the addition of a resistor between the collector and the base:

$$r_\mu \equiv \left( \frac{\partial I_B}{\partial V_{CB}} \right)^{-1} = \beta_0 \left( \frac{\partial I_C}{\partial V_{CB}} \right)^{-1} \approx \beta_0 \left( \frac{\partial I_C}{\partial V_{CE}} \right)^{-1} = \beta_0 r_o \quad (1.57)$$

In integrated circuit transistors  $r_\mu$  is even larger than the value given by (1.57) since a considerable fraction of the base current is due to the reverse injection into the emitter.

The effect of the base-emitter and base-collector junction capacitors can be modeled by adding two extra capacitors called  $C_{je}$  and  $C_{jc}$ , respectively, to the hybrid- $\pi$  model. These capacitors have a value equal to the value of the junction voltage dependent capacitor at the point of operation.

The effect of minority carrier charge in the base can also be modeled using capacitors. A capacitor known as diffusion capacitor can be defined as the derivative of the base charge with respect to  $V_{BE}$ , i.e.,

$$C_b \equiv \frac{\partial Q_F}{\partial V_{BE}} = \frac{\partial (I_C \tau_F)}{\partial V_{BE}} = g_m \tau_F \quad (1.58)$$

This capacitor is connected between the base and the emitter and is thus in parallel with  $C_{je}$ . The parallel combination of  $C_{je}$  and  $C_b$  is usually shown as  $C_\pi$ . A capacitor similar to  $C_b$  can be defined between the collector and the base to account for the minority charge associated with the collector-base minority charge. The value of this capacitor can be calculated to be  $\tau_F/r_o$ . However, in the forward active region this capacitor is much smaller than  $C_{je}$  and hence  $C_\mu$  is dominated by  $C_{je}$ . The intrinsic hybrid- $\pi$  model of the bipolar transistor including the junction and diffusion capacitors is shown in Figure 1.49.

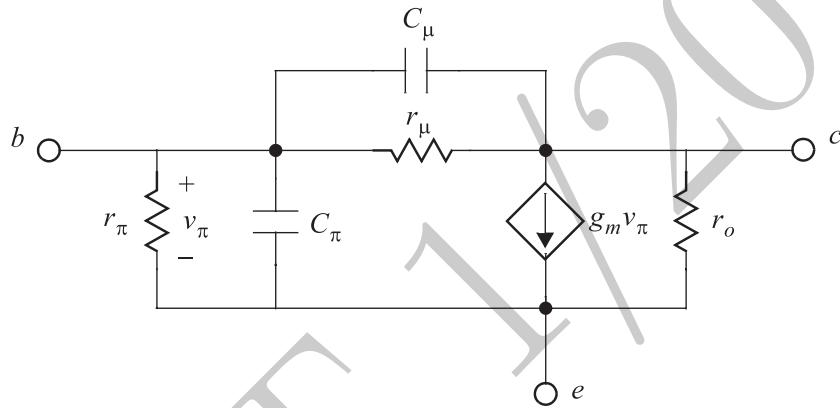


Figure 1.49: The hybrid- $\pi$  model including  $C_\mu$  and  $r_\mu$ .

### Cut-off Frequency, $f_T$

It is customary to quantify the high frequency behavior of a transistor with its *cut-off* frequency,  $f_T$ . Cut-off frequency is the frequency at which the small-signal current gain of the transistor drops to unity due to the parasitic capacitors  $C_\pi$  and  $C_\mu$ . We can easily calculate this frequency using the equivalent circuit for the bipolar transistor in Figure 1.50 .

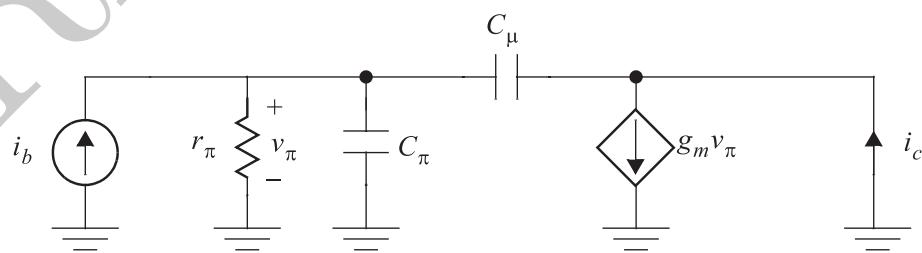


Figure 1.50: The hybrid- $\pi$  model for cut-off frequency calculation. The collector is connected to a constant dc voltage and is hence an ac ground.

The output current can be best measured if the collector is biased using a voltage source. This source is the same as ground in the small signal model and therefore the equivalent circuit will have the following form ( $r_\mu$  is ignored because it is in parallel with  $r_\pi$  which is several orders of magnitude smaller).  $v_\pi$  is related to  $i_b$  as

$$i_b = v_\pi \cdot \left[ \frac{1}{r_\pi} + (C_\pi + C_\mu)s \right] \quad (1.59)$$

The feedforward current through  $C_\mu$  is very small compared to  $g_m v_\pi$ . Ignoring this current, the current gain can be written as

$$\beta(s) = \frac{i_c}{i_b} \approx \frac{g_m v_\pi}{\left[ \frac{g_m}{\beta_0} + (C_\pi + C_\mu)s \right] v_\pi} = \frac{\beta_0}{1 + \beta_0 \frac{C_\pi + C_\mu}{g_m} s} \quad (1.60)$$

A Bode plot of this transfer function is shown in Figure 1.51.

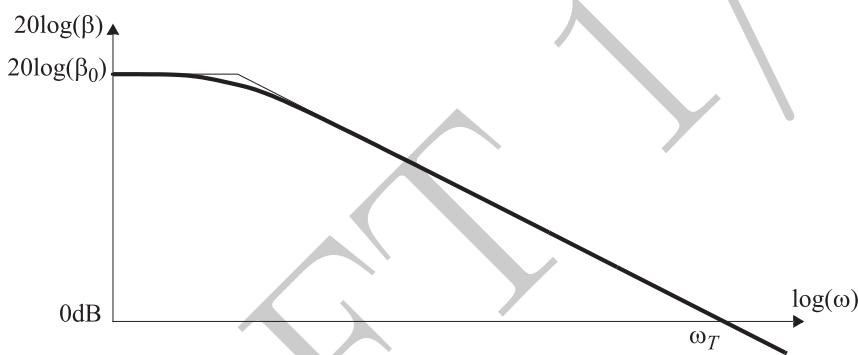


Figure 1.51:  $\beta$  as a function of frequency.

The cut-off frequency,  $f_T$ , can be easily calculated from (1.60) to be

$$\omega_T \equiv 2\pi f_T = \frac{g_m}{C_\pi + C_\mu} \quad (1.61)$$

Now, let us evaluate the cut-off frequency as a function of the collector current. At very low collector current levels,  $g_m$  which is proportional to  $I_C$  is small, and hence  $C_\pi$  is dominated by the junction capacitance,  $C_{je}$ . Since neither junction capacitors have a strong dependence on the collector current, the denominator of (1.61) will have a weak dependence on  $I_C$  and hence the cut-off frequency  $f_T$  will grow linearly with  $g_m$  and so does  $I_C$ . As  $I_C$  is increased though, the base charge capacitance,  $C_b$  becomes comparable and eventually greater than the junction capacitors and thus dominates the denominator of 1.61. Considering that  $C_b = g_m \tau_F$ , assuming that  $\tau_F$  does not have any current dependence (which will see is not true shortly), the cut-off frequency should plateau at  $f_T = 1/\tau_F$ .

However, in practice  $\tau_F$  does increase with the collector current,  $I_C$ , due to the so-called space-charge effect (also known as Kirk effect). This can be explained by noting that when a larger current flows through the base to the collector, there will be more electrons at any given time in transit in the collector which due to their negative electric charge repel other electrons in the base and impede their movement to the collector. At low collector currents, this effect is negligible compared to the tremendous thermal velocity of the electrons, but as the negative space-charge of the collector increases, it becomes more pronounced, making the transit of electrons harder and hence increasing their average transit time,  $\tau_F$ . The end result is that  $f_T$  initially increases with  $I_C$ , reaches a maximum, and drops again, as shown in Figure 1.52.

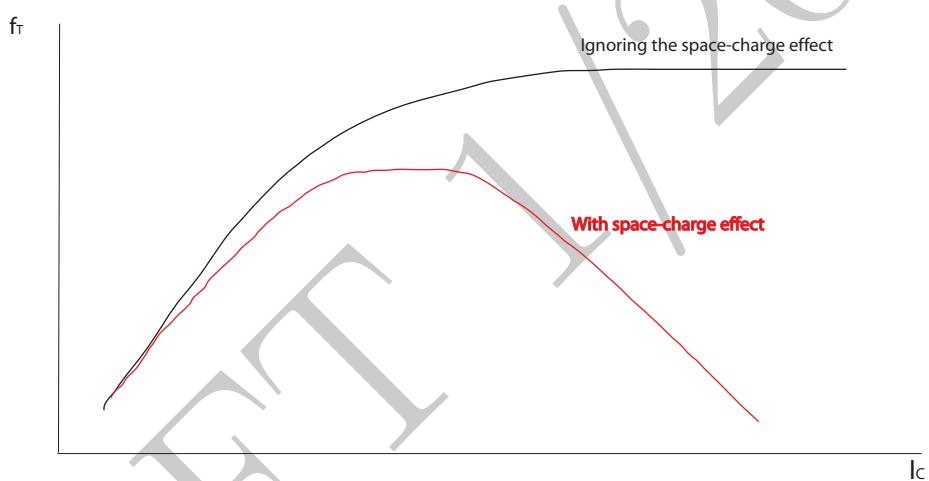


Figure 1.52: The cut-off frequency,  $f_T$ , as a function of the collector current,  $I_C$ .

## 1.4 Metal Oxide Semiconductor (MOS) Capacitor

Online YouTube lectures:

[108N. MOS Capacitor: Energy band diagram, accumulation, depletion, and inversion, threshold voltage](#)

The physics of MOS-capacitors plays an essential role in the analysis of MOS transistors and therefore we will consider MOS-capacitors first. A MOS-capacitor is made of three different materials, a gate made of a conductive material (also referred to as metal, although it is often made of highly doped poly-silicon,) an insulator (usually referred to as oxide, although it could be made out of other materials such as silicon nitride), and a semiconductor bulk.

A simple cross-section of this structure is shown in Figure 1.53.

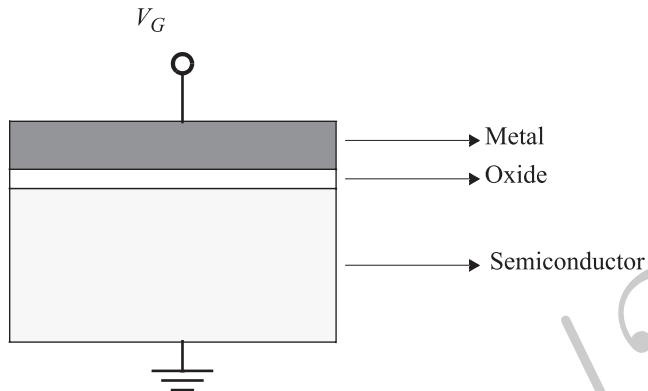


Figure 1.53: A MOS capacitor.

For an *n*-type semiconductor, the energy band diagrams for these three materials in isolation are illustrated in Figure 1.54.  $\Phi_M$  is the metal work

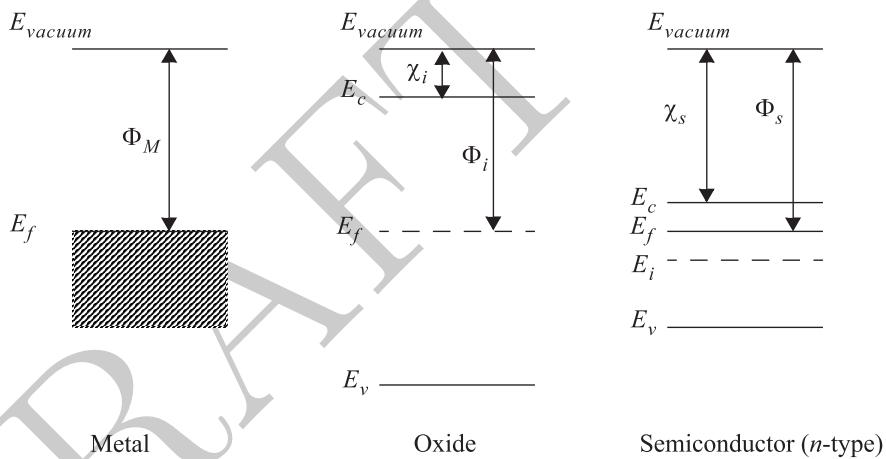


Figure 1.54: Energy band diagrams of a metal (conductor), oxide (dielectric), and semiconductor in isolation.

*function* and is the average amount of energy required to release an electron from the metal into vacuum (shown with the vacuum energy level,  $E_{vacuum}$ ). It can be thought of as the bonding energy of the electron to the metal.  $\chi_i$  is the energy difference between the edge of the conduction band of the insulator and the vacuum energy and is known as the *electron affinity* of the material. In a similar way,  $\chi_s$  is the energy difference between the semiconductor's conduction band

and the vacuum level. We use electron affinity for the semiconductor and the insulator because it does not depend on relative location of the Fermi level and consequently the doping level of the semiconductor material and is determined mainly by the material itself. The work function for the semiconductor, however, is related to the electron affinity by  $\Phi_S = \chi_S + (E_c - E_f)$  and is therefore a function of the doping levels.

Now we bring these three different materials together to form a MOS-capacitor. In doing so we will make two simplify assumptions that will help us illustrate the main points and will later discuss their effects. These two assumptions are the following:

1. There are no charge centers in the oxide or at the oxide-semiconductor interface.
2. Flat band condition exists (same work function for the metal and the semiconductor), i.e.,  $\Phi_M = \Phi_S = \chi_S + (E_c - E_f)$ .

For  $V_G = 0$  the energy band diagram is shown in Figure 1.55. This condition

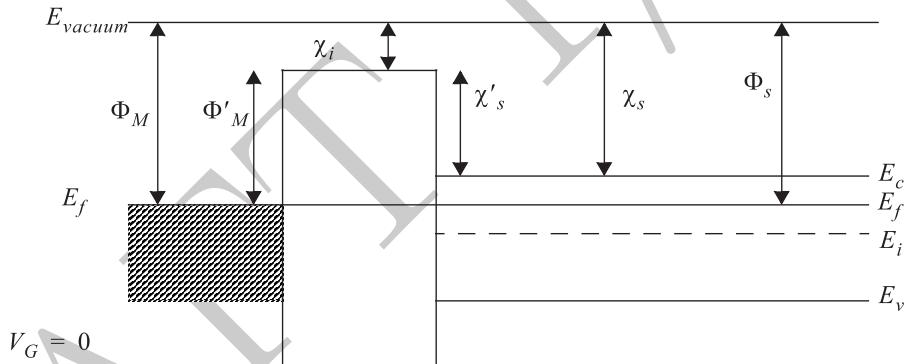


Figure 1.55: Flatband condition when  $\Phi_M = \Phi_S$ .

is known as flat-band for apparent reasons. Applying a non-zero voltage  $V_G$ , will change the energy band diagram. Note that since no dc current can flow in this structure in steady-state, the Fermi level stays constant within the semiconductor. The Fermi level is also pinned at the edge of the conduction band of the metal and will stay constant in the metal. Therefore, the gate voltage will simply change the relative heights of the Fermi level in the semiconductor and the metal, i.e.,

$$E_{f,M} - E_{f,S} = -qV_G \quad (1.62)$$

where  $-q$  is the charge of the electron. First, let us consider the case of  $V_G > 0$ . When a positive voltage is applied to the gate, we expect positive charge to accumulate on the metal at the metal-oxide interface (due to Gauss's law the electric field deep inside a conductor should be zero and hence the charge can

only accumulate at the interface.) Also we expect the density of the electrons to rise inside the semiconductor close to the oxide-semiconductor interface. This charge will be equal to the charge on the gate except for a sign. Now let us see if our intuition agrees with the band diagram picture shown in Figure 1.56.

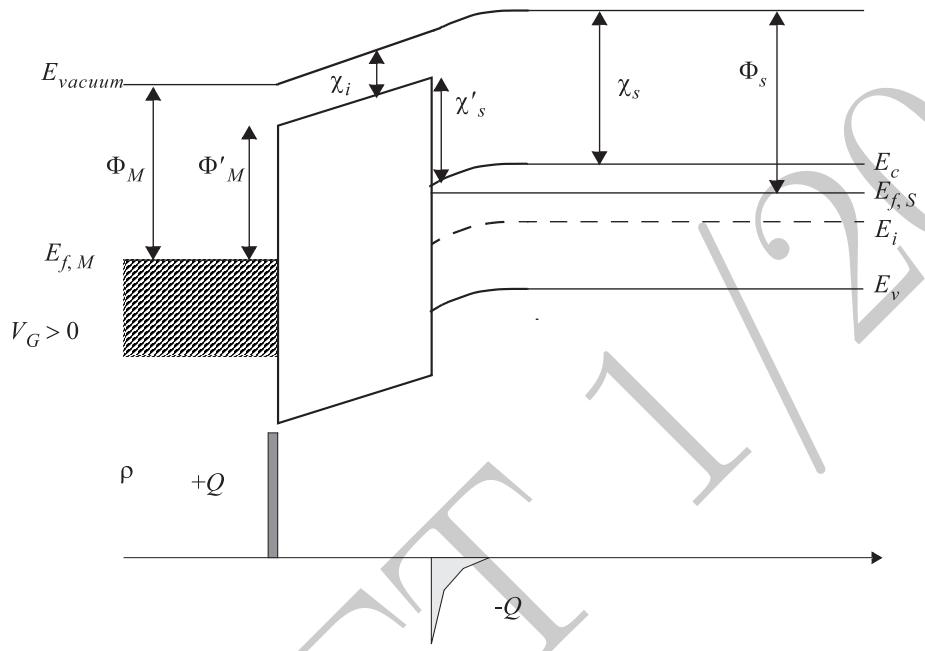


Figure 1.56: Charge accumulation at the surface occurs for  $V_G < 0$  for an *n*-type bulk (or  $V_G > 0$  for a *p*-type one).

As can be seen in the band diagram, the Fermi level gets farther from  $E_i$  and closer to  $E_c$  in the vicinity of the oxide-semiconductor interface. Remembering that the density of the electrons in the semiconductor is given by  $n = n_i e^{(E_f - E_i)/kT}$  (Equation 1.17), we can see that there is an accumulation of majority carriers (in this case electrons) near the interface. For this reason this region of operation is referred to as *accumulation*.

The second case is when we apply a small negative gate voltage. In this case, negative charge will accumulate on the metal at the metal-oxide interface and we need an equal amount of positive charge in the semiconductor to balance it. This will be first achieved by forcing the electrons away from the interface and forming a depletion region consisting of positive ions close to the interface. The band diagram and the charge distribution in this case can be seen in Figure 1.57.

As can be seen from the band diagram, the spacing between  $E_i$  and  $E_f$  is small near the interface which means that a depletion region is formed close to the interface. This region of operation is usually referred to as *depletion*.

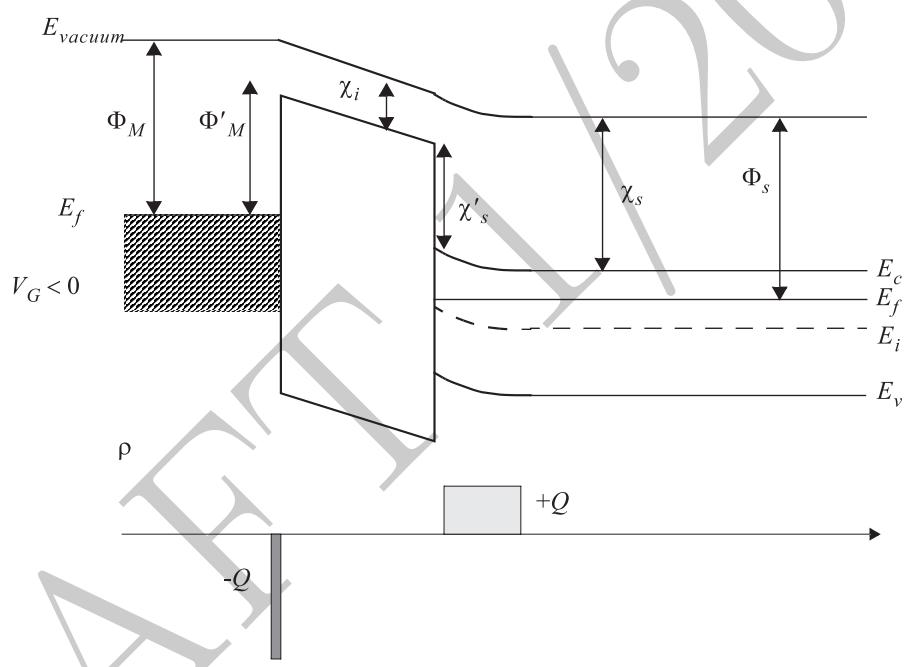


Figure 1.57: Semiconductor's depletion at the surface occurs for  $V_G > 0$  for an *n*-type bulk (or  $V_G < 0$  for a *p*-type one).

If the negative bias of the gate is further increased,  $E_i$  and  $E_f$  will intercept at the semiconductor-oxide boundary for some value of  $V_G$ . Decreasing the gate voltage below this value will result in a band diagram in which  $E_f$  will be below  $E_i$  at the semiconductor-oxide interface. This means that the density of the holes will become larger than the electrons and the surface will become effectively a *p*-type material. This mode of operation is called *inversion* since the type of the surface semiconductor inverts. For small energy differences between  $E_i$  and  $E_f$  the number of induced holes will be very small and therefore, we call this part of the inversion region, *weak inversion*.

As the gate bias is further reduced, there will be a gate voltage where  $E_f$  will be below  $E_i$ , *by the same amount* that  $E_f$  is above  $E_i$  in the bulk of the semiconductor material, as depicted in Figure 1.58. At this point the density

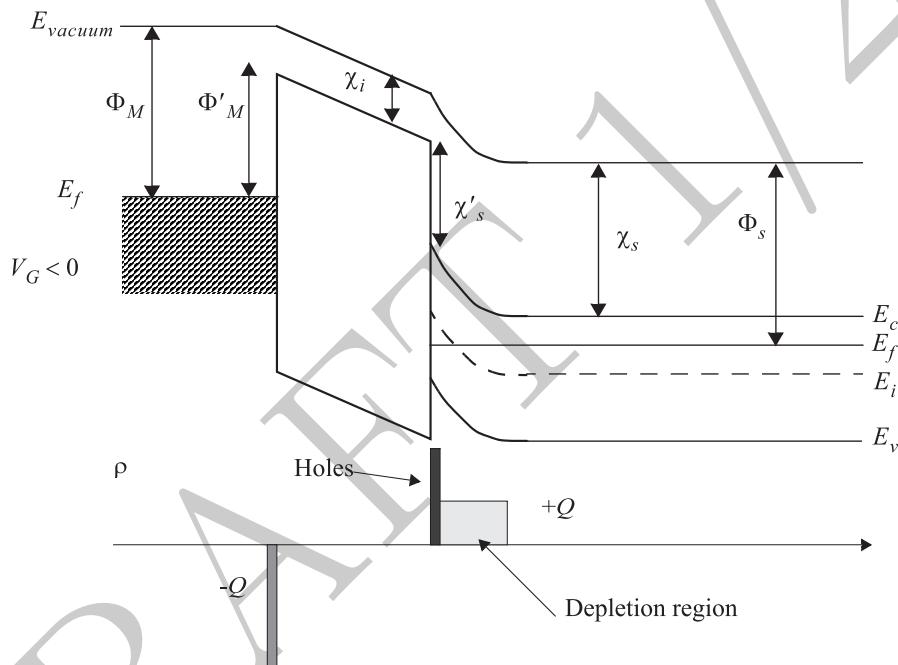


Figure 1.58: Surface inversion.

of the holes at the semiconductor-oxide interface will be equal to the density of the electrons in the bulk of the semiconductor. We define this point as the onset of strong inversion. Note that this definition is rather arbitrary as the onset of strong inversion will depend on the density of the free carriers in the bulk of the semiconductor.

It is noteworthy that in steady state the depletion region does not grow much further for large negative gate bias, rather a thin layer of holes will be formed close to the interface to equalize the charge on the semiconductor side

with the charge on the metal side. In an isolated MOS capacitor, these holes have to come from the generation-recombination process in the bulk of the semiconductors and are therefore slow in formation.

To find the minimum gate voltage for which the strong inversion occurs, we go through two steps: 1. determine the relation between the surface potential  $\phi_s$  and  $V_G$ , 2. determine the surface potential at the verge of strong inversion.

Let us find the relation between  $V_G$  and  $\phi_s$ . We know that

$$V_G = \phi_s + \Delta\phi_{ox} \quad (1.63)$$

where  $\Delta\phi_{ox}$  is the potential drop across the gate oxide. Assuming a uniform doping concentration in the semiconductor, the electric field in the depletion region is given by

$$E(x) = \frac{qN_D}{\epsilon_S}(x - x_d) \quad \text{for } 0 < x < x_d \quad (1.64)$$

where  $\epsilon_S$  is the permittivity of silicon,  $x_d$  is the width of the depletion region, and  $x$  is the distance from the oxide-semiconductor interface. Integrating the above expression, we obtain an expression for the electric potential in the semiconductor as a function of  $x$ , namely,

$$\phi(x) = - \int_{x_d}^x E(x) dx = \frac{qN_D}{2\epsilon_S}(x - x_d)^2 \quad (1.65)$$

and therefore the electric potential at the surface is given by

$$\phi_s = \frac{qN_D}{2\epsilon_S}x_d^2 \quad (1.66)$$

because the electric displacement vector,  $\mathbf{D}$ , should be continuous at the oxide-semiconductor interface,

$$\epsilon_S \cdot E_s = \epsilon_{ox} \cdot E_{ox} \quad (1.67)$$

where  $\epsilon_S$  and  $\epsilon_{ox}$  are the electric permittivity of the semiconductor and the oxide, and  $E_s$  and  $E_{ox}$  are the electric field values at the interface inside the semiconductor and the oxide, respectively. Based on our earlier assumption that there are no charge centers inside the oxide or at the interfaces, the electric field inside the oxide will stay constant at  $E_{ox}$  and therefore the electric potential drop across the oxide will be given by

$$\Delta\phi_{ox} = \frac{\epsilon_S}{\epsilon_{ox}} E_s \cdot t_{ox} \quad (1.68)$$

where  $t_{ox}$  is the thickness of the oxide. Using (1.64) and (1.68), the gate voltage can be expressed in terms of the electric potential at the oxide-semiconductor interface:

$$V_G = \phi_s + \frac{\epsilon_S}{\epsilon_{ox}} t_{ox} \sqrt{\frac{2qN_D}{\epsilon_S} \phi_s} = \phi_s + \gamma \sqrt{\phi_s} \quad (1.69)$$

where  $\gamma$  is defined as:

$$\gamma = \frac{\sqrt{2q\epsilon_S N_D}}{C_{ox}} \quad (1.70)$$

where  $C_{ox}$  is the parallel-plate capacitance per unit area associated with the oxide.

To determine the onset of strong inversion, we notice that the band diagram at the verge of strong inversion that the electric potential at the semiconductor-oxide interface has to be

$$\phi_s = 2\phi_f = 2\frac{kT}{q} \ln \left( \frac{N_D}{n_i} \right) \quad (1.71)$$

where  $q\phi_f = E_f - E_i$  is the energy difference between the intrinsic ( $E_i$ ) and actual ( $E_f$ ) Fermi levels in the bulk of the semiconductor, and  $N_D$  is the donor density in the semiconductor. Keep in mind that  $q\phi_f$  can be calculated from (1.17) considering that at room temperature we are going to be in the middle region of Figure 1.19. Note that further reduction of the gate voltage will not significantly affect the length of the depletion region and will mainly increase the density of the holes at the interface.

The onset of strong inversion is defined by (1.71), therefore, the *threshold voltage* of this idealized structure is given by

$$V'_{T0} = 2\phi_f + \gamma\sqrt{2\phi_f} \quad (1.72)$$

This threshold voltage is identified with the prime sign to emphasize the two simplifying assumption behind it, namely, there are no charge centers inside or at the interfaces of the oxide, and the work functions of the metal and the semiconductor are equal. Neither of these assumptions are true in an actual MOS structure.

If the work functions of the semiconductor and the metal are not equal, the band structure will not be flat for  $V_G = 0$ . Rather we need to apply a gate voltage equal to the difference between the work functions of the metal and semiconductors to reach to the state that the band structure is flat again. Therefore, (1.72) will be offset by

$$\phi_{ms} = \frac{1}{q}(\Phi_M - \Phi_S) = \frac{1}{q}(\Phi_M - \chi_s - E_c + E_f) \quad (1.73)$$

Also there is always a surface charge density at the semiconductor-oxide interface. There may also be a charge density inside the oxide due to impurities and inside the semiconductor due to threshold-adjust implanted dopants<sup>17</sup>. These charge densities will also affect the flat-band voltage, which is the required  $V_G$  for the band diagram to be flat everywhere. This voltage is hence called the *flat-band voltage* and is given by

$$V_{FB} = \phi_{ms} - \frac{Q_{ss}}{C_{ox}} - \frac{Q_i}{C_{ox}} \quad (1.74)$$

---

<sup>17</sup>These are extra dopant atoms introduced by ion implantation to adjust the final threshold voltage to the desired value.

where  $Q_{ss}$  is the surface charge density and  $Q_i$  is the surface charge density of the extra implanted dopant atoms. There may be other terms contributing to the flat-band voltage but the bottom line is that there always exists a gate voltage that results in the flat-band condition.

The threshold voltage given by (1.72) will be shifted by the flat-band voltage, therefore, the threshold voltage of a MOS capacitor is given by

$$V_{T0} = V_{FB} + 2\phi_f + \gamma\sqrt{2\phi_f} \quad (1.75)$$

A very important quantity in a MOS capacitor is the inversion charge at the oxide-semiconductor interface. If we ignore the inversion charge below the threshold voltage, any voltage in excess of the threshold voltage given by (1.75) will linearly increase the inversion charge, with  $C_{ox}$  being the proportionality constant. Therefore, the surface density of the inversion charge can be expressed as

$$Q_{inv} = C_{ox}(V_G - V_T) \quad (1.76)$$

Note that, we simply ignored the free carrier charge in the weak inversion region. As we will see later, this charge is responsible for the subthreshold conduction in MOS transistors.

The ability to modulate the inversion charge using a capacitively isolated terminal (gate) clearly provides an opportunity to make another kind of transistor.

## 1.5 Metal Oxide Semiconductor Field Effect Transistor (MOSFET)

Online YouTube lectures:

[109N. MOSFET Introduction, threshold and body effect, IV characteristic in linear region](#)

Another type of three terminal device is the metal oxide semiconductor (MOS) field effect transistor or MOSFET for short. Consider a MOS-capacitor structure, this time with *p*-type semiconductor, where two pieces of highly doped *n*-type semiconductors have been introduced close to the oxide region, as shown in Figure 1.59. We will call these two regions drain and source for the reason which will become clear shortly.

First assume that the drain and source are both at zero potential ( $V_{DS} = 0$ ). In this case, we have a structure similar to the MOS-capacitor, except for two differences. First, the source and drain *n*+ regions act as a major source of minority carriers (in this case electrons) and therefore the inversion layer can form very fast in contrast with the MOS-capacitor structure in which the minority carriers had to come from the generation recombination process.

The second difference is the existence of the extra voltage source,  $V_{SB}$  which allowing us to change the source-bulk voltage independently of the drain and gate voltage. This extra source will affect the width of the depletion region and hence the threshold voltage defined by (1.75). It is important to note that the

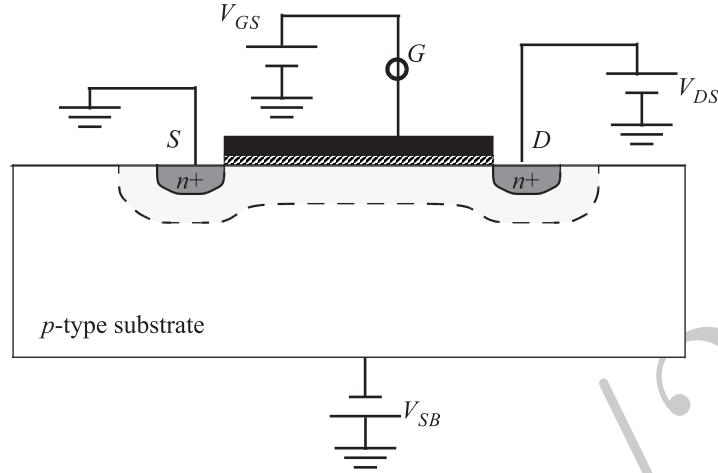


Figure 1.59: Surface inversion.

surface potential has to remain  $2\phi_f$  at the onset of the strong inversion *by definition*. The required electric voltage drop across the oxide for strong inversion is changed from  $\gamma\sqrt{2\phi_f}$  to  $\gamma\sqrt{2\phi_f + V_{SB}}$  to guarantee a surface potential of  $\phi_s = 2\phi_f$  and therefore (1.75) will be modified to

$$V_T = V_{FB} + 2\phi_f + \gamma\sqrt{2\phi_f + V_{SB}} \quad (1.77)$$

which can be rewritten as

$$V_T = V_{T0} + \gamma(\sqrt{2\phi_f + V_{SB}} - \sqrt{2\phi_f}) \quad (1.78)$$

where  $V_{T0}$  is defined in (1.75) and is the threshold voltage for zero source-bulk bias. This dependence of the threshold on the source-bulk bias is known as *body effect* and can play an noticeable role in analog circuits using MOS transistors.

Now we will derive the voltage current relation for a MOS transistor with a channel length,  $L$ , and channel width,  $W$ , as shown in Figure 1.60. Let us first consider a more intuitive picture of the current flow in the transistor. If we assume that  $V_{DS}$  is small, the inversion charge will have a fairly uniform distribution in the channel. In this case, the total charge in the channel will be given by

$$q_{ch} = W \cdot L \cdot Q_{inversion} = W \cdot L \cdot C_{ox}(V_{GS} - V_T) \quad (1.79)$$

The average time it takes for an electron to move from source to drain can be simply approximated as

$$\tau_F = \frac{L}{v_d} = \frac{L}{\mu_n E} = \frac{L}{\mu_n \frac{V_{DS}}{L}} = \frac{L^2}{\mu_n V_{DS}} \quad (1.80)$$

where  $\mu_n$  is the electron mobility in the channel.

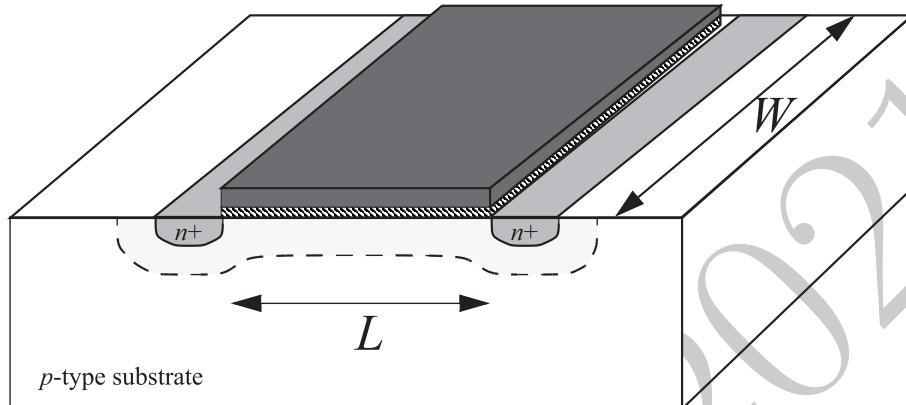


Figure 1.60: Surface inversion.

The drain-source current is simply given by the total charge divided by the time it takes for it to move from the source to the drain, i.e.,

$$I_D = \frac{q_{channel}}{\tau_F} = \frac{WLC_{ox}(V_{GS} - V_T)}{\frac{L^2}{\mu_n V_{DS}}} = \mu_n C_{ox} \frac{W}{L} (V_{GS} - V_T) V_{DS} \quad (1.81)$$

As can be seen for small  $V_{DS}$ , the drain current-voltage behavior is linear and the channel behaves as a linear resistor, whose value is controlled by the gate-source voltage.

The energy band diagram at the surface along the channel gives more intuition about the current conduction. For a small gate voltage, the energy band diagram looks like Figure 1.61. In an NMOS, increasing the gate voltage will

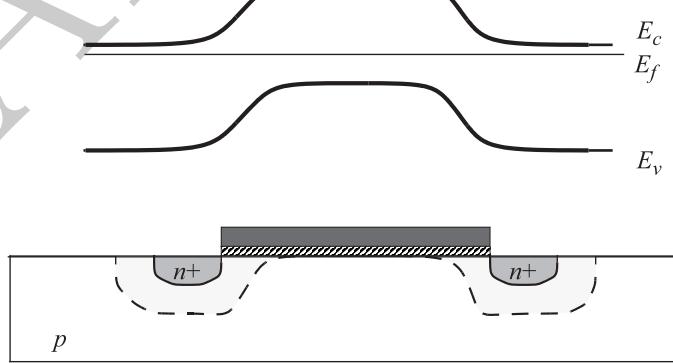


Figure 1.61: Surface inversion.

lower the energy barrier in the channel region. Once above the threshold the

semiconductor material will be  $n$ -type in the channel. Applying a small electric potential difference between the drain and the source, will result in a tilt in the energy band diagram, as shown in Figure 1.62. The electrons will drift through

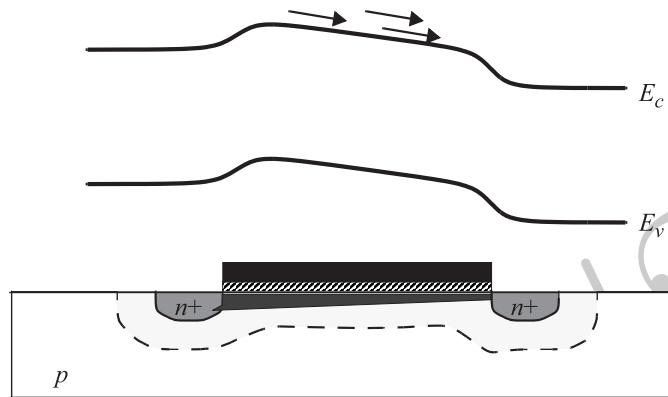


Figure 1.62: Surface inversion.

a continuous piece of  $n$ -type material as the channel is inverted.

Online YouTube lectures:

#### [110N. MOSFET IV Characteristics Derivation, Regions of Operation, Channel length modulation](#)

If  $V_{DS}$  is large, then gate-channel voltage will be smaller at the drain end than the source end and therefore the inversion charge density will be smaller at the drain end compared to the source end. This is shown by drawing the inversion charge channel thicker closer to the source in Figure 1.63.

A more accurate expression for the drain current can be obtained by dividing the channel into infinitesimal MOS transistors in series. This way the  $V_{DS}$  across each transistor is very small, namely,  $dV$  which is the infinitesimal voltage drop across  $dx$ . The incremental inversion charge in this slice is given by (1.76) and is

$$dq \triangleq W \cdot C_{ox} \cdot dx \cdot [V_{GS} - V(x) - V_T] \quad (1.82)$$

where  $V(x)$  is the channel voltage at  $x$ . The drain current,  $I_D$ , of these small MOS transistors is the same as the current through the entire transistor (conservation of charge) and is given by

$$I_D = \frac{dq}{d\tau} = \frac{W \cdot C_{ox} \cdot dx \cdot [V_{GS} - V(x) - V_T]}{\frac{dx^2}{\mu \cdot dV}} = \mu C_{ox} W [V_{GS} - V(x) - V_T] \frac{dV}{dx} \quad (1.83)$$

Note that the drain current of these small transistors are equal to the drain current of the big transistor. Integrating (1.83) across the channel we obtain:

$$I_D \int_0^L dx = \mu C_{ox} W \int_0^{V_{DS}} [V_{GS} - V(x) - V_T] dV \quad (1.84)$$

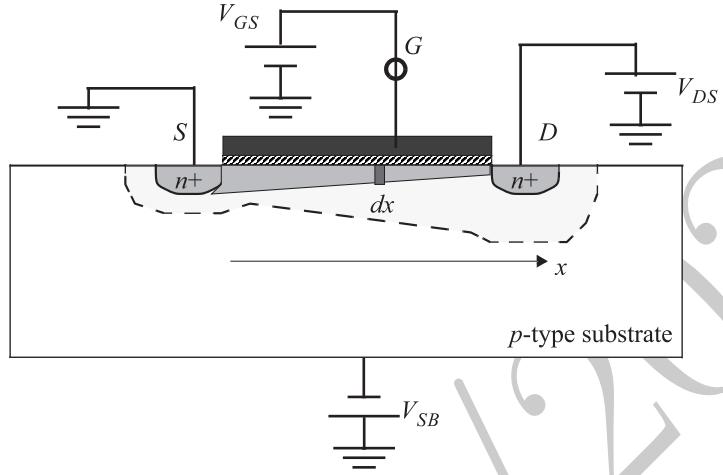


Figure 1.63: Surface inversion.

and therefore

$$I_D = \mu C_{ox} \frac{W}{L} \left[ (V_{GS} - V_T) V_{DS} - \frac{V_{DS}^2}{2} \right] \quad (1.85)$$

There is a simplifying assumption in (1.84) and that is assuming that the  $V_T$  is constant over the channel. However, this is not true as the depletion region underneath the channel is wider below the drain (larger reverse bias between the drain and bulk) compared to the depletion region under the source. It is fairly straightforward to take this effect into account. The theory that accounts for the variable threshold voltage across the channel is known as *bulk-charge theory* and predicts a more accurate  $I-V$  curves for MOS transistors. We will however keep using (1.85) as it is a simpler expression and hence more useful in gaining insight in circuit design.

Note that (1.85) is valid as long as the channel is extended all the way from the source to the drain. The inversion charge in the vicinity of drain is proportional to  $C_{ox}(V_{GS} - V_{DS} - V_T)$ . If  $V_{DS}$  is raised above  $V_{GS} - V_T$ , the inversion charge disappears at the drain, and we enter a different mode of operation known as *pinch-off*, where the channel ends at some point close to the drain but is not extended all the way to the drain, as shown in Figure 1.64.

In pinch-off, the voltage at the edge of the channel will be  $V_{GS} - V_T$  as this voltage defines the onset of the strong inversion at any point in the channel. As  $V_{DS}$  goes above  $V_{GS} - V_T$ , the pinch-off point moves to the left and for a reasonably large  $V_{DS}$  most of the drain-source voltage drops across this relatively short pinch-off region. The electrons are brought to the pinch off point through drift and are attracted into the drain by the high electric field in the pinch-off region. This absorption is similar in nature to absorption of minority carriers into the collector by the electric field. If the channel length,  $L$ , is large compared

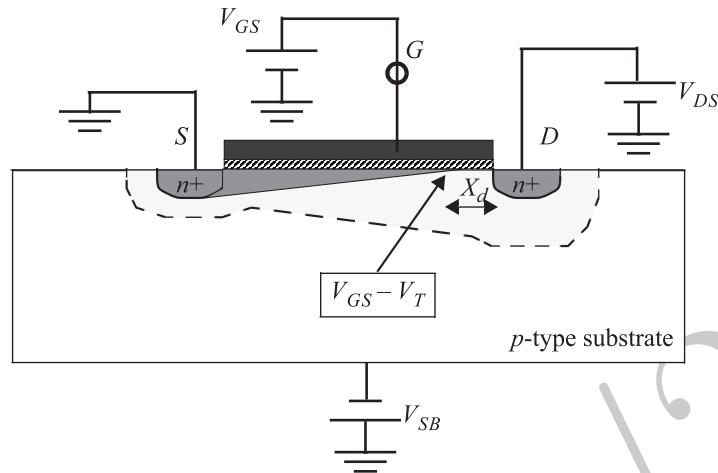


Figure 1.64: Surface inversion.

to the width of the depleted region on the surface,  $X_d$ , the length and the resistance of the channel stays relatively constant. The voltage drop across this effective channel is  $V_{GS} - V_T$  and does not depend on  $V_{DS}$ . Hence, the drain current does not significantly change with  $V_{DS}$  to the first order and will be given by (1.85) with  $V_{DS} = V_{GS} - V_T$ , i.e.,

$$I_D = \frac{\mu C_{ox}}{2} \frac{W}{L} (V_{GS} - V_T)^2 \quad (1.86)$$

which is valid when  $V_{DS} \geq V_{GS} - V_T$ .

Another way of seeing why the current is limited by (1.86) is by looking at the energy band diagram along the channel, as depicted in Figure 1.65. As can be seen, most of the electrons that make it to the edge of the pinch-off region will be absorbed by the electric field, but their number is controlled by the drift in the inverted channel<sup>18</sup>.

As mentioned earlier and can be seen in the above picture, the channel length shrinks with increasing  $V_{DS}$ . This will reduce the effective length of the channel which will in turn increase the current slightly. This effect is known as *channel length modulation* and results in the slant in the  $I_D$  vs.  $V_{DS}$  curves of the MOS transistor, as in Figure 1.66.

The effective channel length can be written as

$$L_{eff} = L - X_d(V_{DS}) \quad (1.87)$$

---

<sup>18</sup>In a way, this is similar to the bipolar junction transistor, in which the number of electrons absorbed by the base-collector electric field is controlled by the number of electrons injected into the base by emitter, which is in turn controlled by the base-emitter voltage.

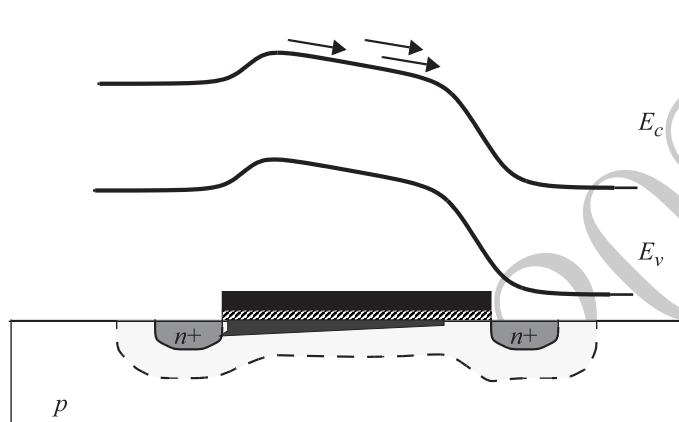


Figure 1.65: Surface inversion.

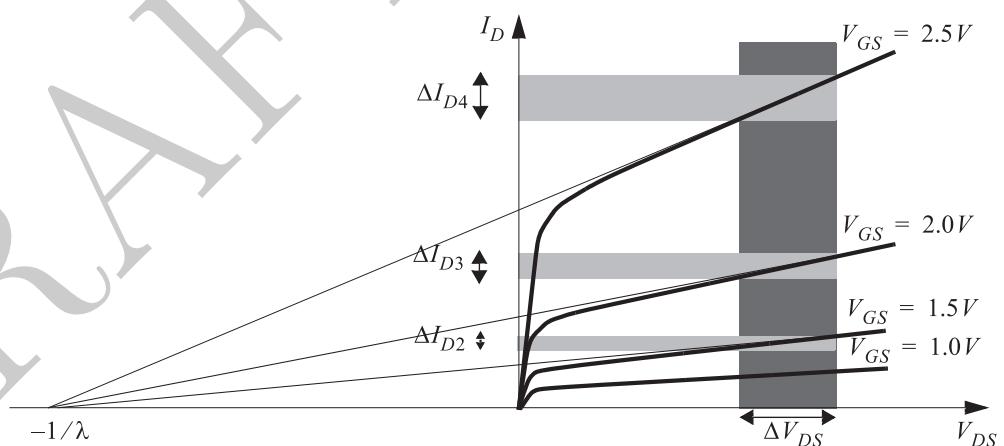


Figure 1.66: Surface inversion.

Therefore, the drain current can be modified as

$$I_D = \frac{\mu C_{ox}}{2} \frac{W}{L_{eff}(V_{DS})} (V_{GS} - V_T)^2 \quad (1.88)$$

The variations in  $I_D$  due to  $V_{DS}$  is given by

$$\frac{\partial I_D}{\partial V_{DS}} = -\frac{\mu C_{ox}}{2} \frac{W}{L_{eff}^2(V_{DS})} (V_{GS} - V_T)^2 \frac{dL_{eff}}{dV_{DS}} = \frac{I_D}{L_{eff}} \frac{dX_d}{dV_{DS}} \quad (1.89)$$

We can define a voltage similar to the Early voltage in the bipolar transistor, given by

$$V_A(L_{eff}) = \frac{I_D}{\frac{\partial I_D}{\partial V_{DS}}} = L_{eff} \left( \frac{dX_d}{dV_{DS}} \right)^{-1} \quad (1.90)$$

In MOS literature it is more common to use the reciprocal of  $V_A$  as the parameter of choice, i.e.,

$$\lambda(L_{eff}) = \frac{1}{V_A(L_{eff})} \quad (1.91)$$

Note the linear relationship between the channel length and the Early voltage. The Early voltage is larger for a larger channel length. This should be intuitively obvious, as the fractional change in the effective channel length will be smaller for a larger  $L$ . The channel length has the same effect as the base width in bipolar junction transistors. However, unlike BJTs, the circuit designer usually has control over the channel length. The drain current can be approximated as

$$I_D = \frac{\mu C_{ox}}{2} \frac{W}{L} (V_{GS} - V_T)^2 [1 + \lambda(L)V_{DS}] \quad (1.92)$$

We will use this expression for the drain current in the pinch-off regime of operation.

Different modes of operation of a MOS transistor are summarized in the following Table<sup>19</sup>:

	$V_{GS} - V_T \geq V_{DS}$	$V_{GS} - V_T < V_{DS}$
$V_{GS} < V_T$	Off $I_D = 0$	Off $I_D = 0$
$V_{GS} \geq V_T$	Triode $I_D = \mu C_{ox} \frac{W}{L} [(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}]$	Pinch-Off $I_D = \frac{\mu C_{ox}}{2} \frac{W}{L} (V_{GS} - V_T)^2 [1 + \lambda(L)V_{DS}]$

<sup>19</sup>To maintain continuity between the triode and pinch-off regions, the triode region equation can also be multiplied by the term  $1 + \lambda(L)V_{DS}$ . This is, however, more useful for simulation purposes and not including it in the analytical expressions does not result in large errors, as the  $V_{DS}$  is generally rather small in triode.

### 1.5.1 Small Dimension Effects

Online YouTube lectures:

- [112N. Velocity saturated MOSFETs, short channel effects, SOI, FinFET, Pillar FET, Strain](#)
- [114N. Kelvin Generator, water FET, positive feedback.](#)

As the MOS transistor is scaled to smaller dimensions, some of the assumptions leading to the developed model may not remain valid. Part of the threshold voltage to form a channel is the voltage used to compensate bulk's depletion charge under the gate oxide. The drain-bulk and source-bulk depletion regions extend below the gate oxide and therefore help with this generation of the depletion region. In a MOS transistor with a large channel length, this effect is negligible, but as the channel length shrinks, these depletion regions become comparable to the bulk depletion region and a smaller gate voltage would be required to deplete and then invert the channel. Thus, the threshold voltage will become smaller as the channel length shrinks if all the other parameters are kept constant, as shown in Figure 1.67. The gate voltage also creates a

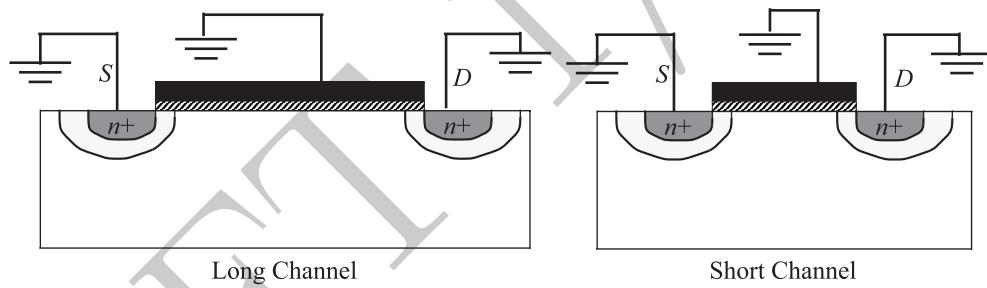


Figure 1.67: Surface inversion.

depletion region, on the sides of the gate electrode due to fringing field. If the channel width,  $W$ , is large, this depletion region constitutes a small fraction of the depletion region under the gate. But if  $W$  becomes small, this fringe depletion region becomes a larger fraction of the depletion region and we need a larger voltage on the gate to maintain this larger depletion region. Therefore, the threshold voltage will increase with decreasing the channel width. This is known as *narrow-width effect* and depicted in Figure 1.68.

In addition to these effects another important change occurs as we go to short channel lengths. Usually the drain voltage and hence the electric field in the channel do not scale as fast as the channel length scaling. If this scaling is not done, the electric fields can reach its critical value and the carrier velocity will saturate as discussed earlier. Velocity saturation will affect the expression for the drain current. If the electric field is much larger than the critical field, the channel charge moves with a constant saturated velocity,  $v_{sat}$ . The total channel charge is given by (1.79), and moves with the velocity  $v_{sat}$ , therefore, it takes  $L/v_{sat}$  seconds for it to go from the drain to the source. Thus the drain

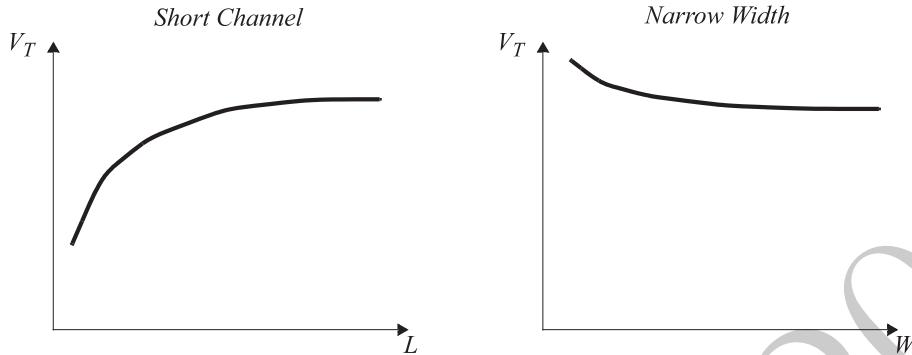


Figure 1.68: Surface inversion.

current is given by the channel charge divided by this time, i.e.,

$$I_D = WC_{ox}(V_{GS} - V_T)v_{sat} = \frac{\mu_n C_{ox}}{2} WE_{sat}(V_{GS} - V_T) \quad (1.93)$$

where \$E\_{sat}\$ is the electric field at which the carrier velocity drops to half the value predicted by constant mobility, i.e.,

$$v_{sat} \equiv \frac{\mu E_{sat}}{2} \quad (1.94)$$

Note that (1.93) is only valid for deeply velocity saturated transistor. A more general expression for the drain current in partially velocity saturated transistor can be obtained using the following simplified expression for the drift velocity, \$v\$,

$$v(E) = \frac{\mu E}{1 + E/E_{sat}} \quad (1.95)$$

Using this expression, and integrating the channel charge over the channel width, we can obtain the following expression for the drain current in saturation

$$I_D = \frac{\mu_n C_{ox}}{2 \left(1 + \frac{V_{GS} - V_T}{E_{sat} L}\right)} \frac{W}{L} (V_{GS} - V_T)^2 \quad (1.96)$$

which reduces to the expressions we have already obtained for the drain current in the long-channel and deep velocity saturated for \$V\_{GS} - V\_T \ll E\_{sat}L\$ and \$V\_{GS} - V\_T \gg E\_{sat}L\$, respectively.

### 1.5.2 Subthreshold Conduction

Online YouTube lectures:

[113N. MOSFET Sub-threshold behavior](#)

As you may have noticed, the definition of the threshold was somewhat arbitrary. We defined the voltage at which the density of the minority inversion charge at the surface becomes equal to the density of the majority carriers in the bulk as the threshold. This by no means indicates that there is no free charge in the channel below the threshold voltage. This free charge in the channel for gate-sources voltages below the threshold results in a drain current known as *subthreshold current*. The value of this current can be estimated by first estimating the inversion charge in the channel. We know that the inversion charge at the surface is given by

$$n_{inv} = n_i \cdot e^{\frac{E_f - E_i}{kT}} \quad (1.97)$$

where  $E_i$  is the intrinsic level at the surface and  $E_f$  is the Fermi level inside the semiconductor.  $E_i$  is controlled by the surface potential,  $\phi_s$ , which is related to  $V_{GS}$  through (1.69). Therefore, the surface charge will be proportional to  $e^{qV_{GS}/nkT}$ , where  $n$  is a factor larger than 1 to take into account the effect of nonlinear dependence of  $V_{GS}$  and  $\phi_s$ . Based on this, the drain current will have an exponential dependence on  $V_{GS}$ .

### 1.5.3 Small-Signal Model for MOS Transistors

Online YouTube lectures:

- 
- [115N. Small-signal model, MOS vs. BJT, core transistor behavior, transconductance](#)  
[116N. Small-signal model, MOS vs. BJT, input and output resistance, capacitance, cut-off](#)

Based on the large signal model derived for the MOS transistor, a small signal model can be derived. Defining transconductance as the derivative of the drain current with respect to the gate-source voltage, in the long channel mode of operation, the transconductance is

$$g_m \equiv \frac{\partial I_D}{\partial V_{GS}} = \mu_n C_{ox} \frac{W}{L} (V_{GS} - V_T) = \sqrt{2\mu_n C_{ox} \frac{W}{L} I_D} \quad (1.98)$$

where  $I_D$  is the dc collector current at the operation point. The transconductance is the slope of the  $I_D$ - $V_{GS}$  curves of the MOS transistor at the operating current of interest, as illustrated in Figure 1.69.

The dc small-signal input resistance at the gate terminal of a MOS transistor is infinite since based on our model, no dc current can pass through gate oxide capacitor. The output resistance of the transistor is similarly give by

$$r_o \equiv \left( \frac{\partial I_D}{\partial V_{DS}} \right)^{-1} = \frac{1}{\lambda(L)I_D} = \frac{L}{I_D} \cdot \left( \frac{dX_d}{dV_{DS}} \right)^{-1} \quad (1.99)$$

which is proportional to the effective channel length,  $L_{eff}$ , to the first order.

The dependence of the threshold voltage on the bulk-source bias results in another current source in the small signal model which has an important effect on the performance of the MOS transistor, when the source is not at a constant

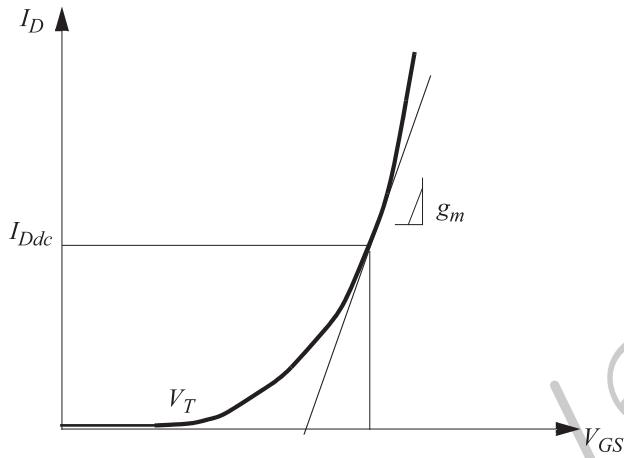


Figure 1.69: Surface inversion.

potential. The changes in the drain current due to body-effect can be quantified using

$$g_{mb} \equiv \frac{\partial I_D}{\partial V_{BS}} = -\mu_n C_{ox} \frac{W}{L} (V_{GS} - V_T) \frac{\partial V_T}{\partial V_{BS}} = \chi g_m \quad (1.100)$$

where

$$\chi \equiv -\frac{\partial V_T}{\partial V_{BS}} = \frac{\gamma}{2\sqrt{2\phi_f + V_{SB}}} = \frac{C_{js}}{C_{ox}} \quad (1.101)$$

where  $C_{js}$  is the capacitance of the depletion region underneath the gate.  $\chi$  is usually between 0.1 and 0.3 in typical CMOS processes. It shows the relative strength of the body effect with respect to the drain modulation due to the gate. The body-effect is sometimes referred to as *back-gate* effect because the bulk behaves as the second gate for the transistor controlling the drain current. Interestingly, the effect of this back-gate is weaker than the effect of the primary gate by the ratio of the depletion capacitance to the oxide capacitance, i.e.,  $\chi$ . The low frequency small signal MOS transistor model is shown in Figure 1.70.

Online YouTube lectures:

### [111N. MOSFET Channel charge in pinch-off, Channel capacitance calculations](#)

To model the dynamic effects in the MOS transistors we need to take different capacitors into account. Both drain and source form reverse junction diodes with the bulk (substrate). As we saw earlier, junction capacitors are nonlinear and

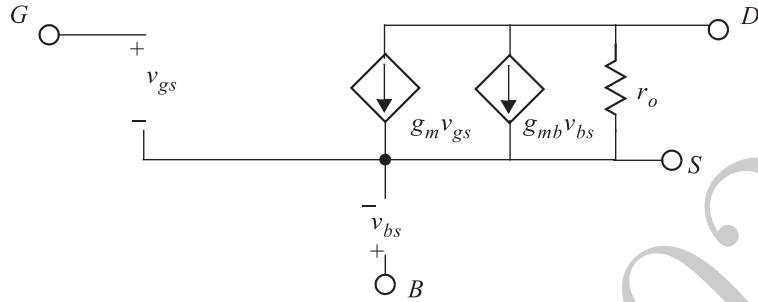


Figure 1.70: Surface inversion.

can be calculated using the following expression:

$$C_{sb} = \frac{C_{sb0}}{\left(1 + \frac{V_{SB}}{\psi_0}\right)^{1/2}} \quad (1.102)$$

$$C_{db} = \frac{C_{db0}}{\left(1 + \frac{V_{DB}}{\psi_0}\right)^{1/2}} \quad (1.103)$$

where  $C_{sb}$  and  $C_{db}$  represent the source-bulk and drain-bulk capacitors, respectively.

In addition to the junction capacitors, the metal-oxide-semiconductor is a capacitor in itself. This capacitor is divided between the source, bulk and drain depending on the mode of operation. These capacitance are shown schematically in the Figure 1.71.

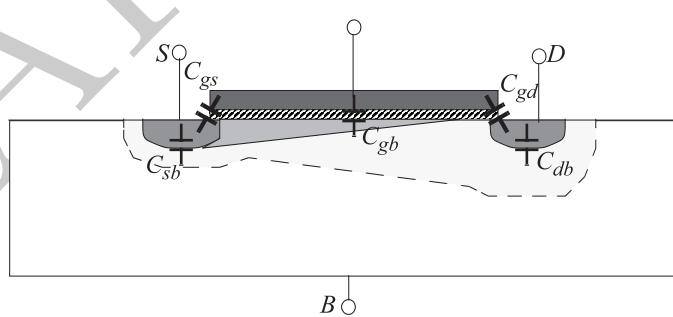


Figure 1.71: Surface inversion.

When the transistor is off, there is no channel and therefore  $C_{gs}$  and  $C_{gd}$  are only limited to the overlap and the fringe capacitor between the gate and the drain and source lateral diffusions, therefore, they are given by  $WL_{ov}C_{ox}$ , where  $L_{ov}$  is the source and drain diffusion extension under the gate. When the

transistor is in triode region, the gate capacitance is equally divided between the drain and source. These two capacitors are therefore given by  $C_{ox}W(L_{ov}+L/2)$ . Once the transistor goes into pinch-off there is no channel charge near the drain and hence  $C_{gd}$  is  $WL_{ov}C_{ox}$ . To calculate  $C_{gs}$  we first have to calculate the channel charge in the pinch-off region. The channel charge in a  $dx$  slice of the channel is given by

$$dq_{ch} = WC_{ox}[V_{GS} - V(x) - V_T]dx \quad (1.104)$$

The total channel charge is given by

$$q_{ch} = WC_{ox} \int_0^{L_{eff}} [V_{GS} - V(x) - V_T]dx \quad (1.105)$$

Using (1.83) we have

$$q_{ch} = \mu_n \frac{W^2 C_{ox}^2}{I_D} \int_0^{V_{GS}-V_T} (V_{GS} - V - V_T)^2 dV = \frac{2}{3} WLC_{ox}(V_{GS} - V_T) \quad (1.106)$$

And therefore channel capacitance is

$$C_{ch} \equiv \frac{\partial q_{channel}}{\partial V_{GS}} = \frac{2}{3} WLC_{ox} \quad (1.107)$$

And hence

$$C_{gs} = W(L_{ov} + \frac{2}{3}L)C_{ox} \quad (1.108)$$

The complete small signal model is illustrated in Figure 1.72.

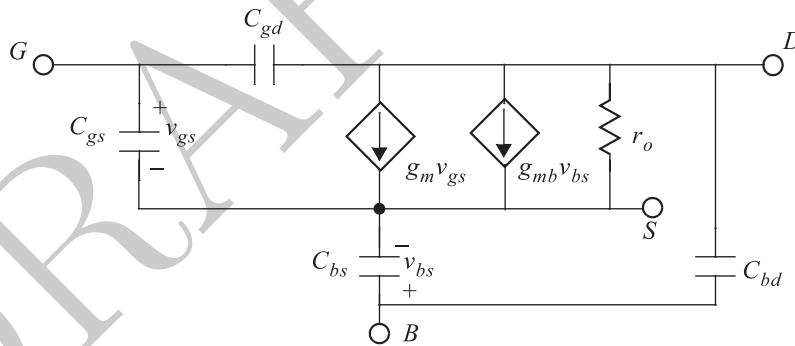


Figure 1.72: Surface inversion.

The *cut-off* frequency,  $f_T$ , of the MOS transistor can be calculated in the same way as the bipolar transistor to be

$$\omega_T \equiv 2\pi f_T = \frac{g_m}{C_{gs} + C_{gd}} \quad (1.109)$$

Ignoring  $C_{ds}$ , using the expression for  $C_{gs}$  in the saturation region,  $f_T$  can be approximated as

$$\omega_T \approx \frac{3}{2} \frac{\mu_n}{L^2} (V_{GS} - V_T) \quad (1.110)$$

As can be seen the cut-off frequency improves with shrinking the  $L$  and having a larger gate overdrive  $V_{GS} - V_T$ . In the velocity-saturated mode of operation, from (1.93) we can calculate the transconductance to be

$$g_m \equiv \frac{\partial I_D}{\partial V_{GS}} = \frac{\mu_n C_{ox}}{2} W E_{sat} \quad (1.111)$$

and the cut-off frequency will be given by

$$\omega_T \approx \frac{3}{4} \frac{\mu_n}{L} E_{sat} \quad (1.112)$$

Note that shrinking the channel length still improves the cut-off frequency but not as fast as the long channel case.

# Chapter 2

## Underlying Mathematics

This chapter is incomplete. For a more detailed review of linear systems and network theory see the [Circuits and System lectures](#) on YouTube.

### 2.1 From Maxwell to Kirchhoff

Online YouTube lectures:

[001. Circuits Fundamentals: Definitions, graph properties, current voltage, power energy](#)

#### 2.1.1 Maxwell Equations<sup>1</sup>

In general the electromagnetic behavior of any system can be determined by solving the Maxwell equations, namely,

$$\nabla \cdot \mathbf{D} = \rho \quad (2.1a)$$

$$\nabla \cdot \mathbf{B} = 0 \quad (2.1b)$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \quad (2.1c)$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \quad (2.1d)$$

where  $\mathbf{E}$  and  $\mathbf{D}$  denote the electric field and electric displacement vectors, respectively<sup>2</sup> and  $\mathbf{H}$  and  $\mathbf{B}$  denote the magnetic field strength and flux density, respectively<sup>3</sup>.

<sup>1</sup>This section technically belongs in the underlying physics chapter, but it makes more sense in the context of the underlying Mathematics.

<sup>2</sup>in a linear isotropic medium  $\mathbf{D} = \epsilon \mathbf{E}$ , where  $\epsilon$  is a scale called the electric permittivity of the medium. For non-isotropic material  $\epsilon$  becomes a tensor, so  $\mathbf{E}$  and  $\mathbf{D}$  may not point in the same direction. In a general non-linear medium  $\mathbf{E} = \mathbf{E}(\mathbf{D})$ .

<sup>3</sup>in a linear isotropic medium  $\mathbf{B} = \mu \mathbf{H}$ , where  $\mu$  is a scale called the permeability of the medium. For non-isotropic material  $\mu$  becomes a tensor, so  $\mathbf{B}$  and  $\mathbf{H}$  may not point in the same direction. In a general non-linear medium  $\mathbf{B} = \mathbf{B}(\mathbf{H})$ .

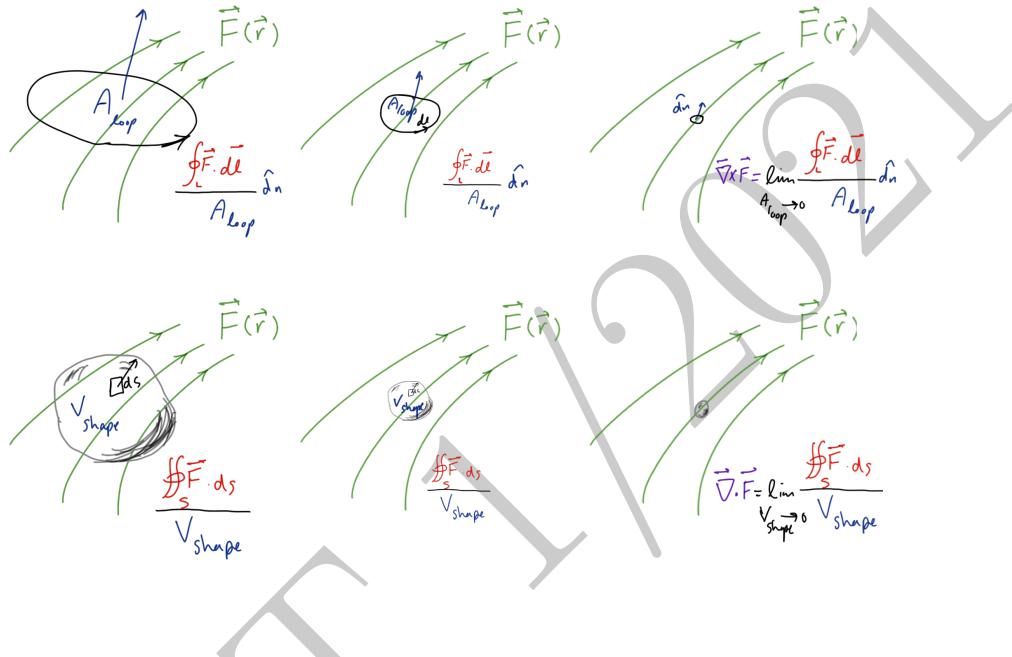


Figure 2.1: Visualization of the concepts of curl and divergence of a vector field.

The divergence of a vector field at a point in space (shown as  $\nabla \cdot \mathbf{F}$  for the filed  $\mathbf{F}$ ) is the volume-normalized flux out of an infinitesimal closed volume ( $v$ ) at that point in space, as shown in Figure 2.1. In other words, it captures the rate that each point in space “bleeds” (or “absorbs”)  $\mathbf{F}$  by taking a small closed volume at that point, calculating the net flux out of it, normalizing it to the volume and taking the limit of the volume to zero to make it a truly localized normalized property of space<sup>4</sup>. This allows us to understand the meaning of the first two equations (Gauss’s law for electricity and magnetism). The first one simply states that the normalized rate space “secrects” electric displacement field,  $\mathbf{D}$ , at any point equal to the electric charge density,  $\rho$ , at that point. Likewise, the second one states that there is no net “bleeding” or “absorbtion” of magnetic flux density,  $\mathbf{B}$ , anywhere; In other words there is no magnetic net charge (no magnetic mono-pole).

The curl of a vector function (shown as  $\nabla \times \mathbf{F}$  for the filed  $\mathbf{F}$ ) is the area-normalized line integral around an infinitesimal closed loop at that point in

<sup>4</sup>This can be mathematically written as:

$$\nabla \cdot \mathbf{F} = \lim_{V \rightarrow 0} \frac{\oint_S \mathbf{F} \cdot d\mathbf{A}}{V}$$

space. Curl is a measure of circular field generation (“twistiness” or, well, “curl”) of the space at each point. It is calculated by taking a small closed loop, calculating the net flux through it, normalizing it to its area and taking the limit of the area going to zero creating a localized normalized property of space<sup>5</sup>.

The third equation, Faraday’s law simply states that the normalized circular component of the electric field,  $\mathbf{E}$  is equal to the (negative) rate of change of the magnetic flux density,  $\mathbf{B}$ . In other words, changing magnetic field produces electric field in a circular fashion. The fourth equation, known as the modified Ampere’s law, states that the normalized circular magnetic field  $\mathbf{H}$  can be generated by either of the current density,  $\mathbf{J}$ , or the changing electric field,  $\mathbf{H}$ , or a combination of them. The last term induced by changing electric field, is sometimes referred to as displacement current, since it behaves like a current density in this equation. These four equations capture the entirety of the classical electrodynamics and are consistent with relativity<sup>6</sup>.

It should also be noted that the conservation of charge is implicit in Maxwell equation. To see this, we can take the divergence of both sides of equation (2.1d), note that divergent of the curl of a vector field is zero<sup>7</sup>, and using (2.1a), we obtain<sup>8</sup>:

$$\nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} = 0 \quad (2.4)$$

which states that the net flux of charge bled from any point in space is equal

<sup>5</sup>This can be mathematically written as:

$$\nabla \times \mathbf{F} = \lim_{A \rightarrow 0} \frac{\oint_L \mathbf{F} \cdot d\mathbf{l}}{A}$$

<sup>6</sup>These equation can also be expressed in the integral form as:

$$\iint_S \mathbf{D} \cdot d\mathbf{A} = q = \int_V \rho, \quad (2.2a)$$

$$\iint_S \mathbf{B} \cdot d\mathbf{A} = 0, \quad (2.2b)$$

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = - \int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{A} \quad (2.2c)$$

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = \mu i_N = \int_S \mathbf{J} \cdot d\mathbf{A} + \int_S \frac{\partial \mathbf{D}}{\partial t} \cdot d\mathbf{A} \quad (2.2d)$$

Stating that the net flux of electrical displacement field,  $\mathbf{D}$ , over a closed surface is equal to the net charge confined within that surface; the net flux of the magnetic flux density,  $\mathbf{D}$ , out of a closed surface is zero; states that the line integral of the electric field,  $\mathbf{E}$ , around a closed loop is inversely proportional to the rate of the change in the net magnetic flux in that loop; and that the integral of the magnetic field,  $\mathbf{H}$ , along a closed loop is given by the net current flowing through the loop plus the rate of the change of the electric displacement field flux through the loop.

<sup>7</sup>Vector calculus identity  $\nabla \cdot (\nabla \times \mathbf{F}) = 0$ .

<sup>8</sup>In integral form it is stated as:

$$\oint_S \mathbf{J} \cdot d\mathbf{A} = \frac{\partial q}{\partial t} = \frac{\partial}{\partial t} \int_V \rho dV \quad (2.3)$$

to the net change in the charge density at that point, *i.e.*, electric charge cannot be created or eliminated<sup>9</sup>.

Another very important corollary of these equation can be seen by taking the curl of (2.1c) and applying (2.1a) and (2.1d) in absence of current or charge ( $\mathbf{J} = 0$  and  $\rho = 0$ ), assuming linear isotropic medium ( $\mathbf{D} = \epsilon\mathbf{E}$  and  $\mathbf{B} = \mu\mathbf{H}$ ) and after a little vector manipulation<sup>10</sup>, we obtain:

$$\nabla^2\mathbf{E} = \epsilon\mu \cdot \frac{\partial^2\mathbf{E}}{\partial t^2} \quad (2.6)$$

which is the three-dimensional wave equation with the wave propagation speed of  $v = 1/\sqrt{\epsilon\mu}$  under these conditions. In vacuum this large yet *finite*, speed of propagation is simply the speed of light<sup>11</sup>.

### 2.1.2 Circuit Approximation

While the general Maxwell equations describe all non-quantum-mechanical electrodynamic systems, there are certain approximations of these equations that lead to simplified cases that can significantly simplify the analysis and lead to additional design insights. One way to create different regions, where more simplified rules apply is comparing the size of the system in question, with the wavelengths associated with the frequencies of interest. This was three regions

<sup>9</sup>All conservation laws of physics have their roots in symmetries in physics. For example, conservation of energy can be derived from the assumption that the laws of physics are the same at different times (homogeneity of time). Similarly, the invariance of laws of physics with respect to space or orientation (homogeneity and isotropicity of space) lead to conservation of linear and angular momentum. Similarly, conservation of charge comes from the fact that adding or subtracting any arbitrary phase to the quantum mechanical wave functions,  $\psi$ , does not change the behavior of the system. For a derivation of conservation law in classical mechanics from symmetries, check the excellent book: *Mechanics: Volume 1 (Course of Theoretical Physics)* by Landau and Lifshitz.

<sup>10</sup>Use the vector calculus identity  $\nabla \times (\nabla \times \mathbf{F}) = \nabla(\nabla \cdot \mathbf{F}) - \nabla^2\mathbf{F}$ . It can be more easily obtained using Einstein index notation for tensors and using the following identity;  $\epsilon_{ijk}\epsilon_{klm} = \delta_{il}\delta_{jm} - \delta_{im}\delta_{jl}$ , where  $\epsilon_{ijk}$  is the three-dimensional Levi-Civita symbol defined by

$$\epsilon_{ijk} = \begin{cases} +1 & \text{if } (i, j, k) \text{ is } (1, 2, 3), (2, 3, 1), \text{ or } (3, 1, 2), \\ -1 & \text{if } (i, j, k) \text{ is } (3, 2, 1), (1, 3, 2), \text{ or } (2, 1, 3), \\ 0 & \text{if } i = j, \text{ or } j = k, \text{ or } k = i \end{cases} \quad (2.5)$$

and  $\delta_{ij}$  is the Kronecker symbol, which is 1 when  $i = j$  and zero otherwise.

<sup>11</sup>One of the highlights of scientific discovery is the well-known story of Maxwell involving the wave equations. Maxwell, wrote the Faraday and original Ampere's equations in a more compact mathematical form and noticed an asymmetry in the equations. He noticed that it could be remedied by adding a term, to Ampere's equation to account allowing for varying electric field to produce circular magnetic field, now known as the displacement current (the last term in (2.1d)). He then noticed that now the interdependence of the electric and magnetic fields can produce sustained waves, where the changing electric field generates changing magnetic field, which in turn produces changing electric field, and so on. He wrote the wave equation and noticed that the speed of the wave is given in free space by  $c = 1/\sqrt{\epsilon_0\mu_0}$ , which were measured independently before. He plugged in the values and got the speed of light! He then postulated that light is electromagnetic wave from purely mathematical observation. A postulate that has had undeniable and long lasting on humanity.

can be defined. When the dimensions of the electrical system in question is much smaller than the wavelength which is called the “circuit approximation” or simply circuits. Another end of the spectrum is when the dimensions are much larger than the wavelengths of interest, which essentially corresponds to optics. The middle range when the dimensions are comparable to the wavelength is sometimes referred to as “distributed circuits” or “microwaves,” although it may or may not correspond to the actual frequencies associated with microwave frequencies. For example, an integrated circuit 1mm on the side, operating at frequency of 10GHz can be considered lumped or in circuit domain, since its length is about 3% of the wavelength and the circuits rules (*e.g.*, KCL and KVL) can be used to model it. However when view at 500THz (corresponding to the color orange), the same chip certainly in the optics domain and can be modeled using geometric optics as a cube since it is much larger than the wavelength of interest (600nm). On the contrary the so-called western interconnect, which is a synchronized 60Hz power distribution network spanning western Mexico, United States, and Canada and is about 5,000km from its farthest edges<sup>12</sup> is in “microwave” domain (distributed), as it is almost exactly one wavelength long.

Another way to view circuit approximation is that the propagation happens instantaneously for lengths of interest as long as they are much smaller than the wavelength. This is similar to saying that the speed of light is infinity ( $c\infty$ ), which ignores radiative and wave effects (such as those predicted by the wave equation (2.6)). This has a couple of interesting and important (and often not explicitly stated) implications. Expressing (2.1d) in linear isotropic medium, in terms of  $\mathbf{E}$  and  $\mathbf{B}$ , as

$$\nabla \times \mathbf{B} = \mu \mathbf{J} + \mu \epsilon \frac{\partial \mathbf{E}}{\partial t}. \quad (2.7)$$

Taking the divergence of both sides and noting that divergence of curl is zero, it can be expressed in terms of speed of light,  $c = \sqrt{(\epsilon\mu)}$ , as:

$$0 = \mu \nabla \cdot \mathbf{J} + \frac{1}{c^2} \frac{\partial \nabla \cdot \mathbf{E}}{\partial t}. \quad (2.8)$$

Now if the propagation is instantaneous for the dimensions of the system ( $c \rightarrow \infty$ ), the last term also goes to zero, leaving  $\nabla \cdot \mathbf{J} = 0$ . This in conjunction with the conservation of charge in (2.4), leads to

$$\frac{\partial \rho}{\partial t} = 0. \quad (2.9)$$

meaning that under these assumption there is not a change in the *net charge* at any point within the circuit. This is the reason that if there are no initial charges on any node in a circuit, *there will be no charge accumulation on any node in a circuit* (which leads to KCL). Also, if there is some *net* charge trapped

---

<sup>12</sup>From somewhere southeast of El Paso, Texas, to three-way border point between Yukon, British Columbia, and Alaska.

on a given node with no dc path out (*e.g.*, in between two capacitor in series), it will stay there.

Instant propagation also means that there cannot be net charge accumulation inside circuit elements and hence for a two terminal element in circuit domain, the instantaneous current going into one terminal is equal to the current leaving the other terminal *at all times*. This is clearly not the case, when the the circuit assumption is broken. For instance, in a half-wavelength dipole antenna there is ac current going into the driving ports of the antenna but there is no current coming out of the tips of the dipole (as evident from the lack of arcs at the tip of a functioning antenna).

### 2.1.3 Element, Node, Loop, Mesh, and Network

A (two-terminal) circuit *element* (also called a branch) is a mathematical abstraction that provides a good approximation for a physical systems with two electrical interface under the instant propagation assumption (size much smaller than wavelengths of interest) discussed earlier. The electrical interfaces are often called terminals or leads. The instant propagation and the resultant no net charge accumulation have important implications that are often used without being stated explicitly (and sometimes without understanding the conditions necessary for their validity).

One is that if the physical system size (the size of the element) is much smaller than the wavelength (instant propagation assumption), the instantaneous current that enters one terminal is *exactly* equal to the current leaving the other terminal *at all times*<sup>13</sup>. While this may sound obvious<sup>14</sup>, it is a non-trivial statement and not true if instant propagation does not hold. For instance, in a  $\lambda/2$  dipole antenna, there is alternating current entering the driving port, but there is no current coming out of the tips of the antenna<sup>15</sup>. This is because the antenna is comparable to the wavelength in dimensions, which means that by the time the signal travel to the tip, the phase of the drive has change significantly due to the finite speed of light that cannot be ignored in this case<sup>16</sup>.

Under the circuit (instant propagation) condition, an electrical system can be modeled as a network of connected circuit elements. The shared connections of two or more leads associated with different elements is called a *node*. Again under circuit conditions, there will be no net charge change of a node<sup>17</sup>. Again, the circuit assumption indicates that all the nodes connected to each other are

---

<sup>13</sup>The second somewhat more subtle one is that it is possible to measure the electric potential of the two terminals, and thus their difference, simultaneously.

<sup>14</sup>Or magical depending on how deeply you think.

<sup>15</sup>As evident from the lack arcs at the tips.

<sup>16</sup>In fact, antenna's ability to efficiently radiates propagating electromagnetic wave while presenting a useful impedance at the port is *EXACTLY* because of this delayed action, also referred to as retarded potential. Understanding this delayed action is the key to understanding antennas and radiation and leads to intuitive, useful, and simply “marvelous” understanding of antenna design, “which this [footnote] is too narrow to contain”.

<sup>17</sup>Or for nodes with a resistive path to ground, no net charge accumulation.

always at the same electrical potential.

Any closed path consisting of elements is called a *loop*. There can be multiple overlapping and nested loops in a circuit. A special kind of loop is a *mesh*, which is a loop that does not contain a smaller loop within itself<sup>18</sup>. The combination of interconnected elements forms a *network*, which is mathematically an interconnected graph with the following interesting relationship between total number of elements,  $e$ , total number of meshes,  $m$ , and total number of nodes,  $n$ ,

$$e = m + n - 1. \quad (2.10)$$

### 2.1.4 KCL and KVL

Online YouTube lectures:

[002. Circuits Fundamental: Passivity and Activity, KCL and KVL, Ideal Sources](#)

#### Kirchhoff Current Law (KCL)

The conservation of charge, with the stronger constraint of no net charge accumulation on any nodes under circuit approximation (dimensions much smaller than the wavelength), lead to Kirchhoff current law. Considering a node in the circuit where multiple elements are connected to (as shown in Figure 2.2), the algebraic (meaning they can be both positive or negative) sum of the charges injected into the node at any point has to be zero:

$$\sum_{n=1}^N q_n(t) = 0 \quad (2.11)$$

This is the result of conservation of charge when there is no charge accumulation on the node. Note that in general electrodynamics it is entirely possible for charge to accumulate on a conductor and change its value over time, invalidating (2.12). For circuit, the currents flowing into the nodes as by definition the rate of change in charge, *i.e.*, the derivative of (2.12),

$$\sum_{n=1}^N i_n(t) = 0 \quad (2.12)$$

which is simply the KCL, which states that the instantaneous algebraic sum of the currents into any given node of a circuit should be zero at all times.

#### Kirchhoff Voltage Law (KVL)

KCL is simply the application of Faraday's law, (2.1c) (or its integral form (2.2c)), which itself is a manifestation of the conservative nature of the electric

<sup>18</sup>Sometimes, the analogy to glass window panes is used to describe meshes.

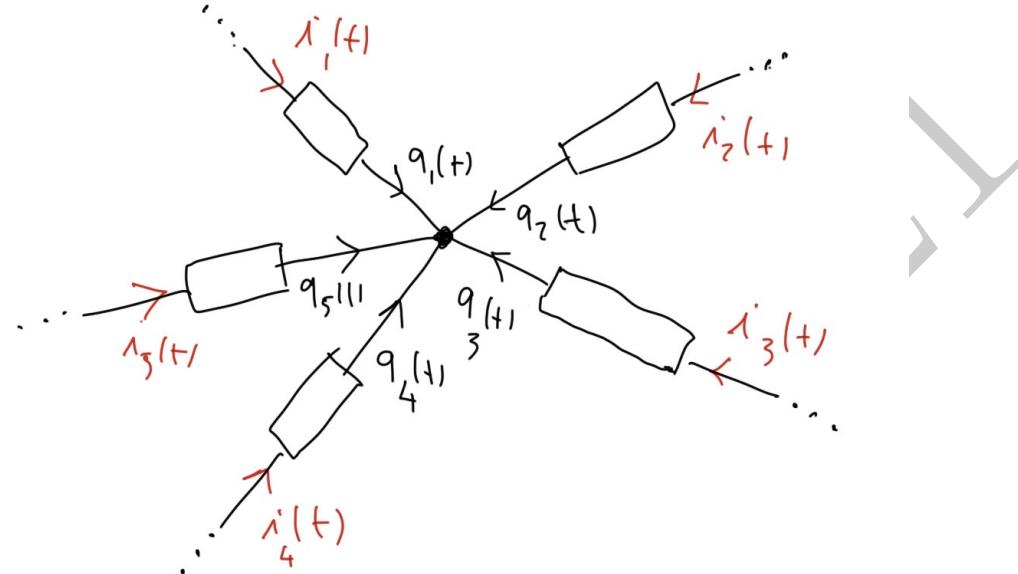


Figure 2.2: Charge and current into a node of the circuit.

field (subject to magnetic induction)<sup>19</sup>. Considering a loop of elements within a circuit (Figure 2.3), equation (2.2c) indicates that the algebraic (meaning they can be both positive or negative) sum of element voltages in the loop is equal to the change of magnetic flux passing through that loop,  $\Phi_B$ , *i.e.*,

$$\sum_{m=1}^M v_m(t) = -\frac{d\Phi_B}{dt} \quad (2.13)$$

In term on the right hand side is important in the study of devices involving magnetic coupling, such as transformers and is an important contributor to errors in the presence of electromagnetic interference (EMI). In the absence of time-varying magnetic coupling from outside, (2.13) reduces to the more conventional statement of KVL, namely,

$$\sum_{m=1}^M v_m(t) = 0. \quad (2.14)$$

It simply states that in the absence of magnetic coupling the algebraic sum of the element voltages in the direction of the loop is zero<sup>20</sup>.

<sup>19</sup>Which is in turn a manifestation of conservation of energy, which in turn comes from homogeneity of time.

<sup>20</sup>The voltages should be counted in the same way, *e.g.*, counting moving along the loops from the positive terminal of each element to its negative terminal as addition and from the

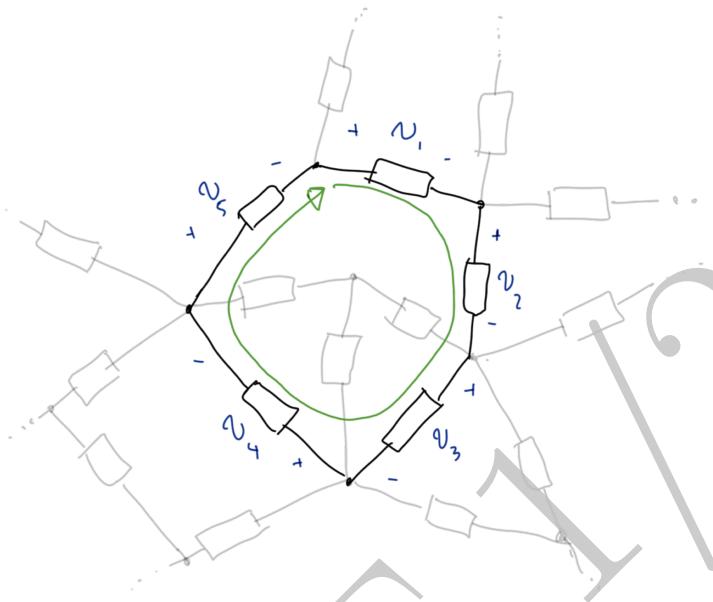


Figure 2.3: Summation of voltages through a loop of elements.

## 2.2 Network Theory

### 2.2.1 Nodal Analysis

Online YouTube lectures:

[004. Ground, Y-Matrix, Node Voltage Stimulus vectors](#)

[005. Examples, Dependent Sources, Existence of a Solution](#)

[006. Dependent Sources, w/ Voltage Sources, Super Nodes](#)

For a shorter review, watch:

[117N. Circuit and network theory brief overview, nodal analysis, theorems.](#)

A network of linear elements consisting of resistors, capacitors, inductors, and dependent and independent current and voltage sources in general can be described in terms of its node voltages by applying the KCL at different nodes of the network. In general voltages should be defined in terms of the electric potential difference between two nodes. For a network with  $N + 1$  nodes, it is customary and often convenient to consider one of the nodes as the reference

---

negative to positive as subtraction.

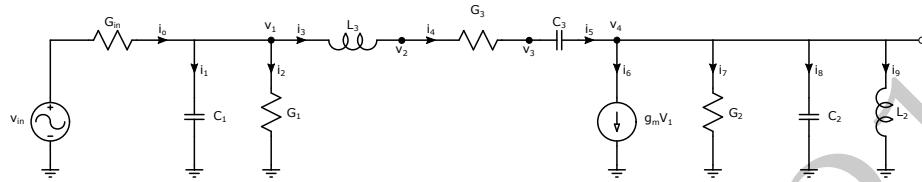


Figure 2.4: Nodal analysis example.

(zero) voltage and call it the ground. This reduces the number of the node voltages to be considered by one to  $N$ .

Note that there is no need to apply the KCL to the reference (ground) node, as the remaining  $N$  nodal equations automatically subsume the KCL applied to the ground node, and hence provide us with  $N$  independent equations which is exactly the number of equations we need to solve for the  $N$  unknown nodal voltages,  $v_1, v_2, \dots, v_N$ .

The nodes of the network are connected to each other via branches that can consist of two terminal elements such as resistors, capacitors, inductors, and dependent and/or independent sources (voltage or current), as shown in the example network of Figure 2.4.

A branch could also consist of series combination of multiple two terminal devices. First, we treat all nodes equally, namely we deal with nodes in the middle of an independent branch not connected to anything else (e.g., nodes 2 and 3 in Figure 2.4) the same way as all other nodes. We can write four independent KCL equations for each of the four non-ground nodes of the circuit (nodes 1 through 4), are respectively:

$$\begin{aligned} -i_0 + i_1 + i_2 + i_3 &= 0 \\ -i_3 + i_4 &= 0 \\ -i_4 + i_5 &= 0 \\ -i_5 + i_6 + i_7 + i_8 + i_9 &= 0 \end{aligned} \tag{2.15}$$

These four KCL equations can be expressed in terms of the node voltages as:

$$\begin{aligned} -G_{in} \cdot (v_{in} - v_1) + (C_1 s + G_1) \cdot v_1 + \frac{1}{L_3 s} \cdot (v_1 - v_2) &= 0 \\ -\frac{1}{L_3 s} \cdot (v_1 - v_2) + G_3 \cdot (v_2 - v_3) &= 0 \\ -G_3 \cdot (v_2 - v_3) + C_3 s (v_3 - v_4) &= 0 \\ -C_3 s \cdot (v_3 - v_4) + g_m \cdot v_1 + (C_2 s + G_2 + \frac{1}{L_2 s}) \cdot v_4 &= 0 \end{aligned} \tag{2.16}$$

Note that since  $v_{in}$  is an input (excitation) node, we should move the term associated with that to the right-hand side of (2.16), if our objective is to find

$v_1$  through  $v_4$  in terms of the excitation. These four equations can be written in the matrix form as,

$$\begin{bmatrix} G_{in} + G_1 + C_1 s + \frac{1}{L_3 s} & \frac{1}{L_3 s} & 0 & 0 \\ -\frac{1}{L_3 s} & \frac{1}{L_3 s} + G_3 & -G_3 & 0 \\ 0 & -G_3 & G_3 + C_3 s & -C_3 s \\ g_m & 0 & -C_3 s & G_2 + (C_2 + C_3)s + \frac{1}{L_2 s} \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} G_{in}v_{in} \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (2.17)$$

which can be written as:

$$\begin{bmatrix} Y_{11}(s) & Y_{12}(s) & Y_{13}(s) & Y_{14}(s) \\ Y_{21}(s) & Y_{22}(s) & Y_{23}(s) & Y_{24}(s) \\ Y_{31}(s) & Y_{32}(s) & Y_{33}(s) & Y_{34}(s) \\ Y_{41}(s) & Y_{42}(s) & Y_{43}(s) & Y_{44}(s) \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} G_{in}v_{in} \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} G_{in} \\ 0 \\ 0 \\ 0 \end{bmatrix} \cdot v_{in} \quad (2.18)$$

This process can be generalized to a circuit with  $N$  independent nodes, in which case, the system of equations describing the system can be written as:

$$\begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1N} \\ Y_{21} & Y_{22} & \cdots & Y_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{N1} & Y_{N2} & \cdots & Y_{NN} \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix} = \begin{bmatrix} i_1^s \\ i_2^s \\ \vdots \\ i_N^s \end{bmatrix} \quad (2.19)$$

This can be written in a more compact fashion as

$$Y \cdot V = I^s \quad (2.20)$$

In the case of a single input  $v_{in}$ , 2.20 can be stated as<sup>21</sup>

$$Y \cdot V = \begin{bmatrix} Y_1^S \\ Y_2^S \\ \vdots \\ Y_N^S \end{bmatrix} \cdot v_{in} = Y^S \cdot v_{in} \quad (2.21)$$

Generally we are trying to solve for the node voltages,  $v_1, v_2, \dots$ , i.e., the vector  $V$ , which can be done by inverting the  $Y$  matrix, i.e.,

$$V = Y^{-1} \cdot I^s \quad (2.22)$$

In the case of a single input single output (SISO) system, usually one of  $v_i$ s is considered the output. Without loss of generality, we can assume that  $v_2$  is the output node. In this case, we can solve for the system transfer function,

<sup>21</sup>In the case of a multiple input system, we cannot factor the input perturbation out of the  $Y^S$  vector. Nonetheless, this does not pose a problem as the right-hand side can be written as a vertical  $I^S$  vector which is a function of various input stimuli (voltage and current), but the rest of the analysis remains similar to the above treatment in nature.

$H(s)$ , using the Cramer's rule:

$$H(s) \equiv \frac{v_2}{v_{in}} = \frac{\det \begin{vmatrix} Y_{11} & Y_1^S & \cdots & Y_{1N} \\ Y_{21} & Y_2^S & \cdots & Y_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{N1} & Y_N^S & \cdots & Y_{NN} \end{vmatrix}}{\det \begin{vmatrix} Y_{11} & Y_{12} & \cdots & Y_{1N} \\ Y_{21} & Y_{22} & \cdots & Y_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{N1} & Y_{N2} & \cdots & Y_{NN} \end{vmatrix}} = \frac{\Delta'}{\Delta} = \frac{N(s)}{D(s)} \quad (2.23)$$

where the numerator,  $N(s)$ , has the same roots as the determinant of the modified  $Y$  matrix in which the second column replaced with the  $Y^S$  vector from (2.21). We show this determinant as  $\Delta' \equiv \det |Y'|$  and the roots of the denominator,  $D(s)$  are the simply the roots of the determinant of the  $Y$  matrix, shown as  $\Delta$ .

A very important corollary can be made here, noting that the poles of the transfer function are determined by the roots of the determinant of the  $Y$  matrix, namely, the complex frequencies at which it is singular (the determinant is zero). As is evident from (2.19),  $\Delta$  is independent of the definition of the input and output nodes, and is purely a function of the circuit elements and their interconnections. The poles, being the natural frequencies of the system, are the same no matter where the excitation is applied and where the output is taken.

On the other hand, zeros that are the roots of the matrix  $Y'$  depend on the choice of the input and output. For instance, if the output is considered to be a node other than  $v_2$  in the above example, the matrix numerator admittance matrix would be a different matrix,  $Y''$  with a different determinant and hence in general different roots. Hence the zeros of the transfer function, unlike its poles, depend on both the network itself and the choices of the input and/or the output.

It should also be kept in mind that any modification to the circuit that changes the determinant of  $Y$  results in a change in the general behavior of the circuit including its natural frequencies.

### Y Matrix: Compact Form

Although writing the nodal equations in the forms of (2.15) and (2.16) was very mechanical and hence very suitable for computer simulation of the circuit, it is possible to write those equations in a simpler form. Suppose we do not particularly care about the voltages of the inner nodes of the middle branch of the circuit (nodes 2 and 3). In this case we note that the second and the third KCL equations in (2.15) are redundant since  $i_3 = i_4 = i_5$  because  $L_3$ ,  $G_3$  and  $C_3$  are all on the same branch and thus have the same current.

It is often more convenient and economical to consider these nodes (e.g.,  $n_2$  and  $n_3$ ) as non-essential nodes and develop a reduced form of the nodal equations

with a lower degree that does not include these node voltages explicitly and only deals with the node voltages at the ends of the explicitly series branches, in this example nodes  $n_1$  and  $n_4$ . Even if the intermediate node voltages were required, this representation is not a major impediment as determining the intermediate voltages is a straightforward voltage divider calculation once  $v_1$  and  $v_4$  are known<sup>22</sup>.

Once the essential nodes are identified, the KCL can be applied at each of these nodes. In this case, the four KCL equations of (2.15) reduce (eliminating  $i_4$  and  $i_5$  from the two middle equations) to

$$\begin{aligned} -i_0 + i_1 + i_2 + i_3 &= 0 \\ -i_3 + i_6 + i_7 + i_8 + i_9 &= 0 \end{aligned}$$

Noting that the current in each branch is simply given by the reactance of the branch times the difference between the voltages of the two nodes in which it is terminated, we can write:

$$i_3 = (G_3 + C_3s + \frac{1}{L_3s}) \cdot (v_1 - v_4) \quad (2.24)$$

Using this we have,

$$\begin{bmatrix} G_{in} + G_1 + G_3 + (C_1 + C_3)s + \frac{1}{L_3s} & -(G_3 + C_3s + \frac{1}{L_3s}) \\ g_m - (G_3 + C_3s + \frac{1}{L_3s}) & G_2 + G_3 + (C_2 + C_3)s + (\frac{1}{L_2s} + \frac{1}{L_3s}) \end{bmatrix} \begin{bmatrix} v_1 \\ v_4 \end{bmatrix} = \begin{bmatrix} G_{in} \\ 0 \end{bmatrix} \cdot v_{in} \quad (2.25)$$

which results in a smaller dimension, denser matrix to represent the circuit. It is easy to verify that the determinant of this compact  $Y$  matrix is the same as the original sparse one.

## 2.2.2 Mesh Analysis

YouTube Lecture:

### [007. Mesh Analysis: Mesh Analysis, 3D Networks, Super Mesh](#)

There is an equally valid, but generally less useful, dual to the nodal analysis, called Mesh analysis. In Mesh analysis the mesh sub-currents are taken as the essential unknowns and the KVL is invoked around various meshes in the circuit, stating that the total voltage drop around the mesh should be zero. This voltage drop can be stated in terms of individual mesh sub-currents and the branch impedance. This results in a matrix relation:

$$Z \cdot I = V^s \quad (2.26)$$

where diagonal elements of  $Z$ , namely,  $Z_{ii}$ s are called *self-impedances* of different meshes and the off-diagonal elements,  $Z_{ij}$ , (when  $i \neq j$ ) are referred to as *mutual-impedances* between meshes,  $i$  and  $j$ .

<sup>22</sup>One a slightly more mathematical note, writing the nodal equations for all the nodes including the non-essential ones results in a larger, yet more sparse  $Y$  matrix with fewer non-zero elements. This matrix can be reduced to a smaller, denser matrix, noticing that the original matrix is not full ranked.

DRAFT 1/2021

## Chapter 3

# Low Frequency Behavior of Transistor Circuits

Economics of integrated circuits dictates a lot of choices in IC and is one of the causes for the different topologies being used in integrated circuits vs. discrete electronics. The following online YouTube lectures provides a very brief summary:

[Economics of Integrated Circuits, Yield, Pricing](#)

A generic understanding and a side-by-side comparison of similar topologies using different technologies is given in two online YouTube lectures:

- 
- (Pt.1) [Amplifier Fundamentals, MOS, BJT, and ATD \(arbitrary 3-terminal device\), maximum gain](#)  
(Pt.2) [Amplifier Fundamentals, MOS, BJT, and ATD \(arbitrary 3-terminal device\), maximum gain](#)

### 3.1 Bipolar Amplifiers

Online YouTube lectures:

[Emitter Degeneration, Transfer characteristics and gain\), maximum gain](#)

#### 3.1.1 Common Emitter Topology

As we discussed earlier, bipolar junction transistor (BJT) converts variations in the base-emitter voltage to changes in the collector current. To make a voltage amplifier we need to convert these variations in current to voltage variations again. This opens up the possibility of using different types of loads, the most obvious one being a single resistor. Common emitter stage is the most commonly used stage for general voltage amplification. It can provide a reasonable voltage and current gain at the same time. The following figure shows the basic common

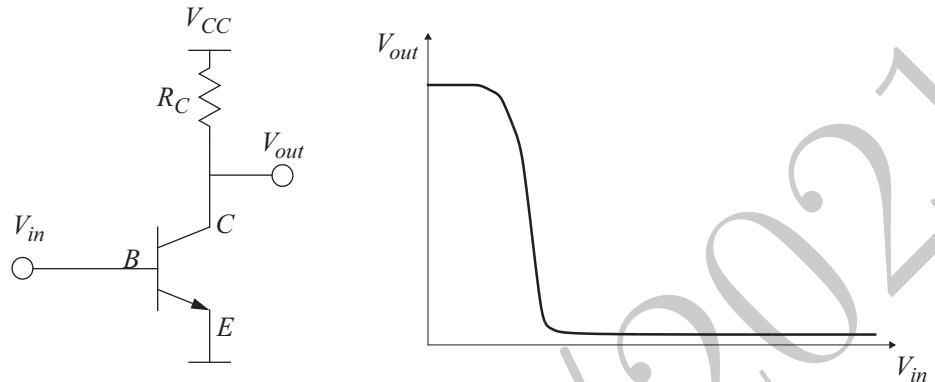


Figure 3.1: The basic common emitter stage with its input output transfer characteristic.

emitter stage. The idea is to use the strong dependence of the collector current on the base-emitter voltage to obtain amplification, as shown in Figure 3.1.

The nonlinear transfer characteristic of the above amplifier can be easily calculated from the collector current equation of the bipolar transistor in forward active region, i.e.,

$$I_C = I_S \cdot \exp\left(\frac{V_{BE}}{V_T}\right) \quad (3.1)$$

where  $I_S$  is the junction saturation current and  $V_T = kT/q$  is the thermal voltage. The output voltage is given by the supply voltage minus the voltage drop across the load resistor,  $R_C$ :

$$V_{out} = V_{CC} - R_C I_S \exp\left(\frac{V_{in}}{V_T}\right) \quad (3.2)$$

which is valid as long as the transistor does not enter saturation, i.e.,  $V_{out}$  is larger than  $V_{sat}$  (0.1-0.2V).

A powerful tool in understanding the behavior of the amplifiers with transistors in them is the load line. Consider the  $I_C$ - $V_{CE}$  curves for the bipolar transistor again. These curves represent the collector currents vs. collector-emitter voltage for different base-emitter voltages. The curves will not have equal spacing for equal steps in the base-emitter voltage due to the exponential behavior of the transistor, as seen in Figure 3.2.

Another constraint between the current and voltage is imposed by the resistor. The current and the voltage of the resistor are related through Ohm's law, i.e.,

$$V_{CE} = V_{CC} - R_C I_C = V_{CC} - V_R \quad (3.3)$$

where  $V_R = R_C I_C$  is the voltage drop across the resistor. Equation (3.3) can be represented with a straight line in the above  $I_C$ - $V_{CE}$  plot. Changing the input,

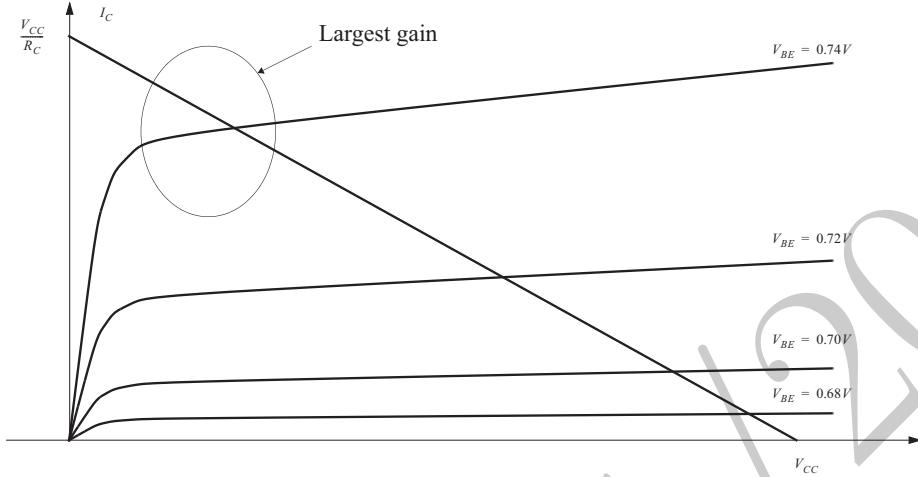


Figure 3.2:  $I_C$  vs.  $V_{CE}$  curves of a BJT together with a resistive load line.

changes the associated  $I_C$ - $V_{CE}$  plot of the system. The output voltage and the collector current will be given by the intersection of the transistor curves and the resistor load line. As can be seen from this graph the maximum gain can be achieved when the spacing between the  $I_C$ - $V_{CE}$  curves is maximum since a given change in the input voltage will result in the largest change in the output voltage. The small signal gain of the amplifier can be calculated by differentiating (3.2):

$$A_v \equiv \frac{\partial V_{out}}{\partial V_{in}} = -\frac{R_C}{V_T} I_S \exp\left(\frac{V_{in}}{V_T}\right) = -R_C \cdot \frac{I_C}{V_T} \quad (3.4)$$

which confirms that ac gain is proportional to the collector current and therefore the maximum gain is achieved at the highest collector current. The small signal voltage gain can be more easily calculated from the small signal model introduced earlier. The gain can be calculated using the simplified, low-frequency  $\pi$ -model shown in Figure 3.3.

If the output resistance of the transistor is large, compared to  $R_C$  the gain is simply given by

$$A_v = -g_m R_C \quad (3.5)$$

Keeping in mind that  $g_m = I_C/V_T$ , it can be seen that we have obtained the same result as (3.4) from the small signal model more easily. Equations (3.4) and (3.5) may lead one to think that increasing  $R_C$  we can increase the gain without bound. As  $R_C$  is increased,  $r_o$  cannot be ignored anymore and has to be taken into account. In this case the gain will be modified to

$$A_v = -g_m (R_C \parallel r_o) \quad (3.6)$$

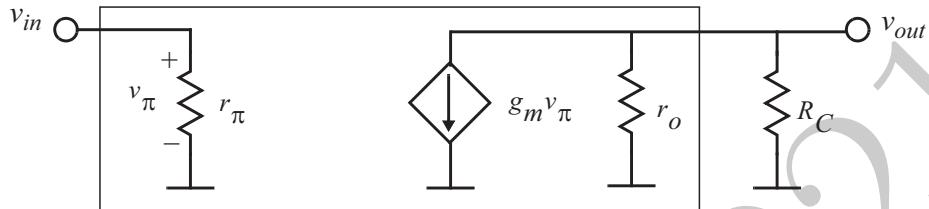


Figure 3.3: The small signal model for the common emitter stage of Figure 3.1.

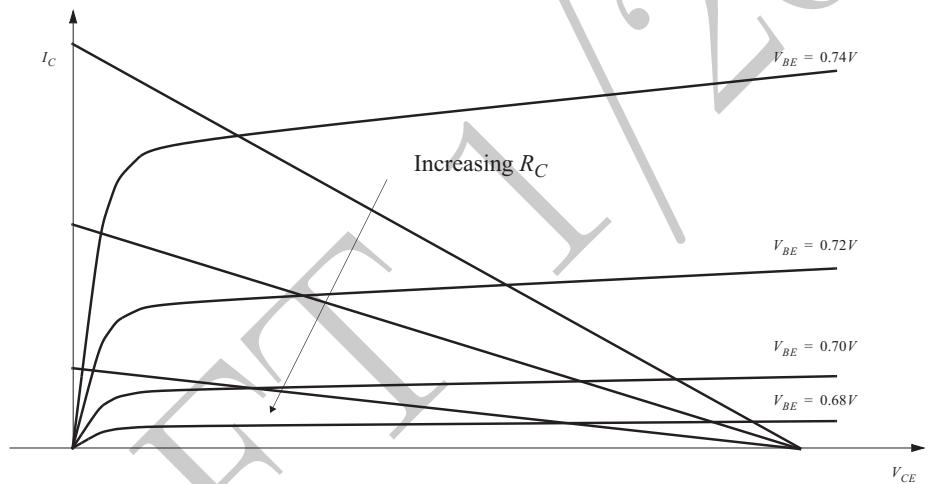


Figure 3.4: The effect of increasing  $R_C$  on the load line.

Therefore, the limit of the achievable low-frequency gain from the basic common-emitter stage is

$$|A_{max}| = g_m r_o = \frac{I_C}{V_T} \cdot \frac{V_A}{I_C} \quad (3.7)$$

In modern bipolar process technologies, this ratio is between 200 and 2000.

The effect of increasing the load resistor can be better understood by using the load line concept. Increasing the load resistor is equivalent to a reduction in the slope of the load line. Therefore a given change in the base-emitter voltage will result in a larger change in the output voltage (or equivalently,  $V_{CE}$ ). This can be seen in Figure 3.4.

In the limit, the load line will be horizontal and the gain will be limited by the output resistance of the transistor. This can be visualized using the picture shown in Figure 3.5.

A small change in the input voltage will result in a change in the collector

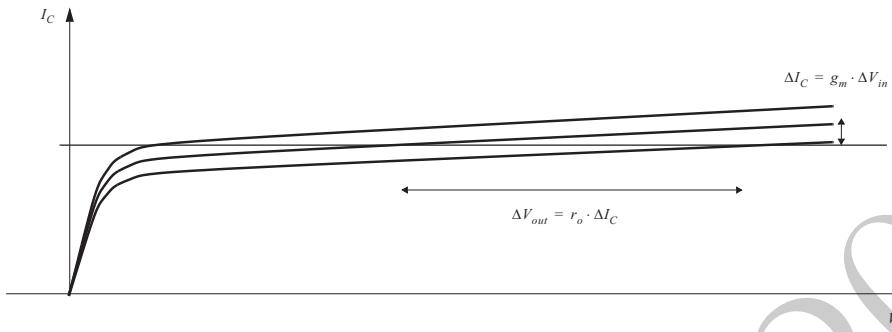


Figure 3.5: The load line of a current source.

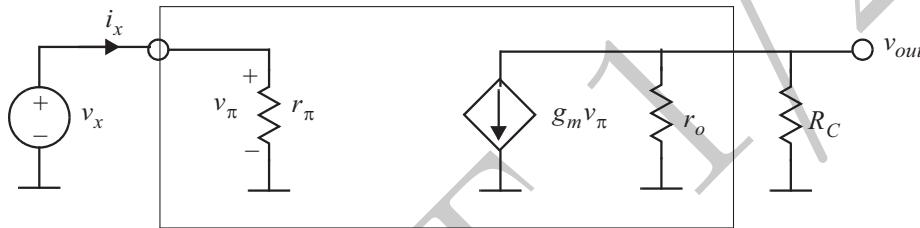


Figure 3.6: The input impedance of Common-Emitter Stage.

current, which is  $g_m$  times larger. This small change in the collector current, however, can only be obtained by a large change in the collector-emitter voltage. This change will be proportional to the change in the collector current with a proportionality constant given by  $r_o$ .

Other important parameters of an amplifier are the input and output impedance of the circuit. The input and output impedances can be calculated by setting the independent voltage sources to zero, applying a test voltage (or current) source to the port of interest and finding the current (or voltage) flowing in the source. The ratio of the test voltage to the test current will provide the port resistance. As a simple and illustrative example, let us consider the simple common-emitter amplifier. The input impedance is calculated by small signal test voltage,  $v_x$ , to the input (base) and measuring the current,  $i_x$ , as illustrated in Figure 3.6.

In this case it is clear by inspection that the ratio of the test voltage to the test current is given by  $r_\pi$ . Therefore,

$$R_{in} \equiv \frac{v_x}{i_x} = r_\pi \quad (3.8)$$

The output impedance can be calculated in the same fashion by applying a test current or voltage source to the output and setting all the independent

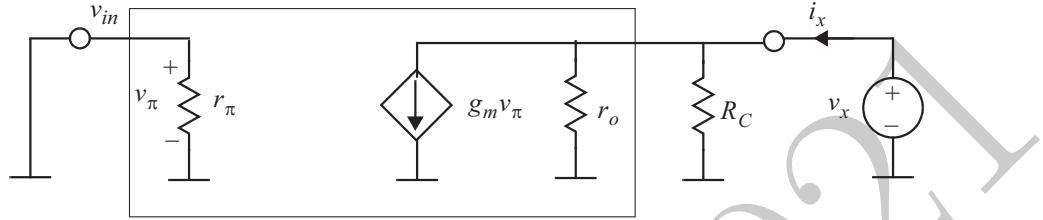


Figure 3.7: The output impedance of Common-Emitter Stage.

sources (including the input voltage source) to zero. This will set  $v_\pi$  to zero and hence the dependent current source to zero, as shown in Figure 3.7.

The output resistance is given by

$$R_{out} \equiv \frac{v_x}{i_x} = r_o \parallel R_C \quad (3.9)$$

Although there was no need for the test current and voltage sources in this case and both the input and output impedance of the basic common emitter stage could be obtained by inspection, in general, the test current and voltage source are very useful tools in determining of the input and output impedance of more complicated topologies.

### 3.1.2 Emitter Degeneration

As can be seen from the transfer function for the basic common-emitter amplifier. Its transfer function is very nonlinear and therefore results in a lot of distortion in the input signal. This is due to the limited useful range of input voltage, which is partly because of the high-gain in the stage and partly because the input voltage is directly dropping across the base-emitter junction and is directly affected by the highly nonlinear transfer characteristic of BJTs. This effect can be reduced by allowing only a fraction of the input voltage to drop across the base-emitter junction. The most effective way of attaining this voltage division is to use a resistor in the emitter as shown in Figure 3.8.

Here we introduce the small signal T-model to analyze the behavior of the emitter degenerated common-emitter amplifier. The T-model can be obtained from the  $\pi$ -model through the steps depicted in Figure 3.9.

Note that  $r_m = 1/g_m$  is the effective small signal resistance looking into the emitter.

Ignoring  $r_o$  for the time being, the small signal model for the emitter follower is shown in Figure 3.10.

As can be seen the current in the series combination of arm and RE is solely controlled by the input voltage (as you have a voltage source in parallel with a current source), i.e.,

$$i_e = \frac{v_{in}}{R_E + \alpha r_m} \quad (3.10)$$

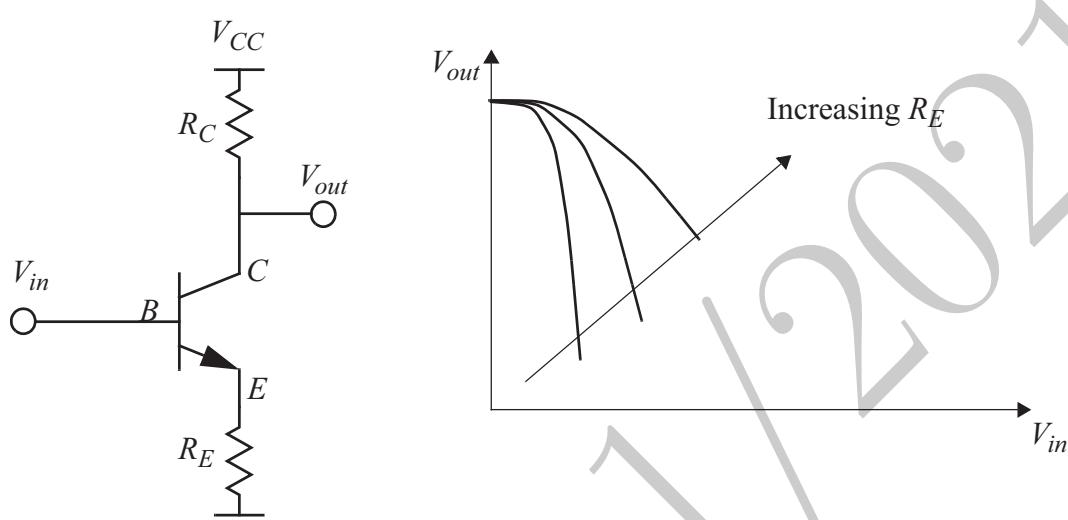


Figure 3.8: The common-emitter with emitter degeneration.

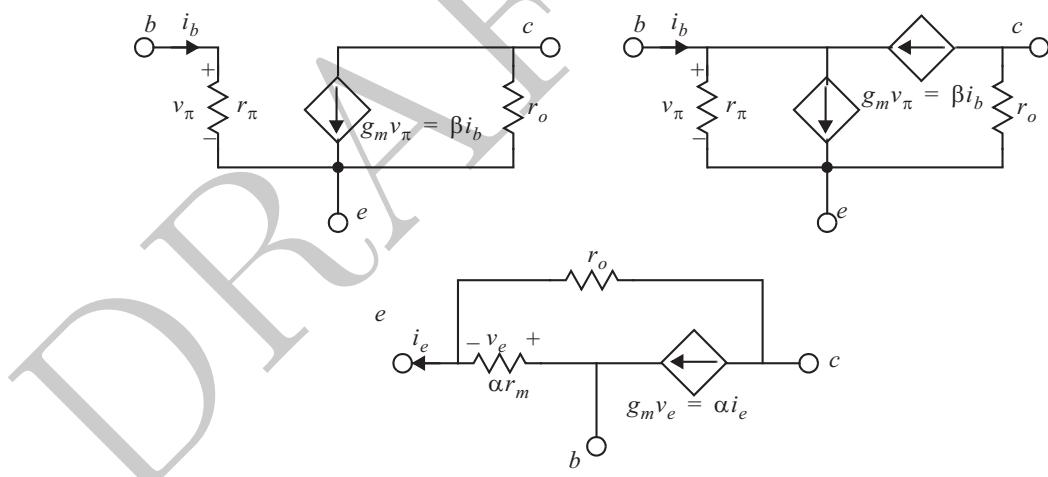


Figure 3.9: The conversion of the T-model to the  $\pi$ -model.

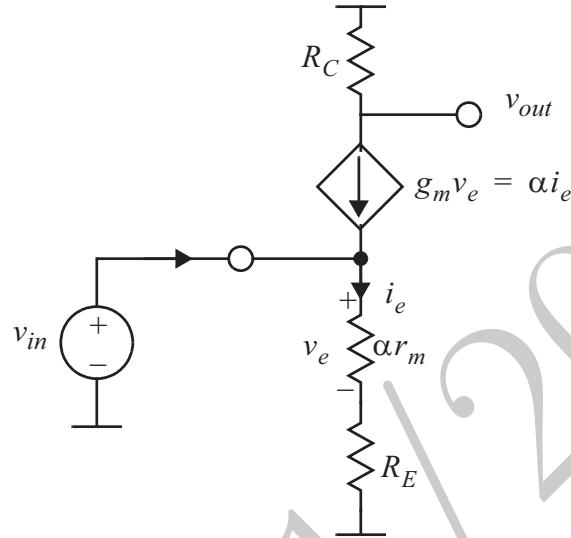


Figure 3.10: The small-signal model for the emitter degenerated common emitter.

Also it is easy to note that the output voltage is controlled by the dependent current source and the load resistor, i.e.,

$$v_{out} = -\alpha R_C i_e \quad (3.11)$$

Therefore, the voltage gain is given by

$$A_v = \frac{v_{out}}{v_{in}} = -\frac{\alpha R_C}{R_E + \alpha r_m} = -\frac{g_m R_C}{1 + g_m R_E / \alpha} \equiv -\frac{g_m R_C}{1 + g_m R_E} \quad (3.12)$$

Now looking carefully at the small signal model and (3.12), we can make a very important and useful observation. Noting that the emitter current appears in the collector with a gain of  $\alpha$  (which is usually very close to one) and also noting that the emitter current is controlled by the total resistance in the emitter, the gain of the common-emitter stage in general will be  $\alpha$  times the total impedance in the collector divided by the total impedance in the emitter, i.e.,

$$A_{CE} = -\alpha \cdot \frac{R_{C,\text{total}}}{R_{E,\text{total}}} \quad (3.13)$$

Hence as long as the emitter resistor is considerably larger than  $\alpha r_m$  the gain will be determined by the ratio of the collector to the emitter resistors. This is a very important property as it will desensitize the gain of the stage to the transistor nonlinear behavior and allows us to control it by adjusting the resistor values. In fact, due to this lack of dependence on the transistor characteristics,

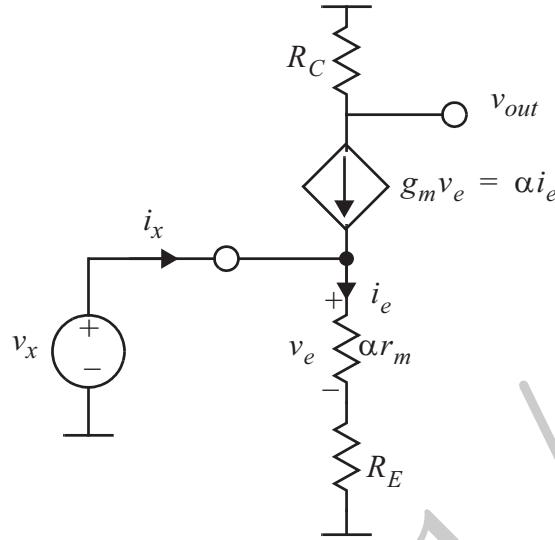


Figure 3.11: The small-signal model for the input impedance of the emitter-degenerated common emitter.

the bipolar transistor can be replaced by any other similar three-terminal device (such as MOSFET or JFET) and still yield a similar gain.

The input impedance can be calculated by applying a test current source and calculating the current as shown in Figure 3.11.

The current flowing in the emitter is simply given by (4.10). Therefore, the current through the test voltage source is

$$i_x = (1 - \alpha)i_e = \frac{i_e}{\beta + 1} = \frac{1}{\beta + 1} \cdot \frac{v_x}{R_E + \alpha r_m} \quad (3.14)$$

Thus, the input resistance will be

$$R_{in} \equiv \frac{v_x}{i_x} = (1 + \beta) \cdot (R_E + \alpha r_m) = r_\pi + (1 + \beta)R_E \approx r_\pi + \beta R_E \quad (3.15)$$

As can be seen, the impedance in the emitter is multiplied by a factor  $(1 + \beta)$  when seen from the base. This should not be very surprising as the base current is smaller by the same factor.

The output resistance of the equivalent circuit shown in the previous page is simply  $R_C$ . This is a good approximation as long as  $R_C$  is considerably smaller than  $r_o$  (which is usually the case). But it is instructive to find the output resistance in the presence of  $r_o$  to gain more insight into the behavior of the output resistance. To do this we apply a test voltage source at the output and set the input to a constant dc voltage or equivalently ac ground, as illustrated in Figure 3.12.

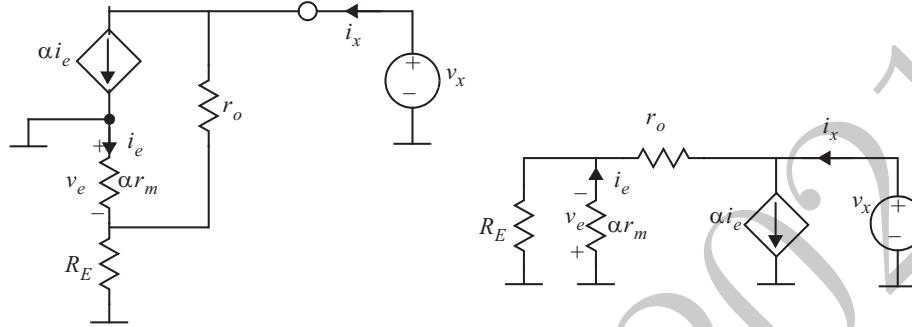


Figure 3.12: The small-signal model for the output impedance of the emitter-degenerated common emitter.

Since  $r_o$  is much larger than the parallel combination of  $R_E$  and  $\alpha r_m$ , the current through  $r_o$  is approximately  $v_x/r_o$ . This current is divided between  $R_E$  and  $\alpha r_m$ . Therefore,

$$i_e \approx -\frac{R_E}{\alpha r_m + R_E} \cdot \frac{v_x}{r_o} \quad (3.16)$$

from which we can calculate  $i_x$  to be

$$i_x \approx \frac{v_x}{r_o} + \alpha i_e = \frac{v_x}{r_o} \cdot \left(1 - \frac{\alpha R_E}{\alpha r_m + R_E}\right) \approx \frac{v_x}{r_o} \cdot \left(\frac{r_m + R_E/\beta}{r_m + R_E/\alpha}\right) \quad (3.17)$$

and therefore

$$R_{out} \approx R_C \parallel r_o \cdot \left(\frac{r_m + R_E/\alpha}{r_m + R_E/\beta}\right) \approx R_C \parallel r_o \cdot \left(\frac{1 + g_m R_E}{1 + g_m R_E/\beta}\right) \quad (3.18)$$

For typical values of  $R_E$ ,  $g_m R_E \ll \beta$  and therefore the total output resistance can be approximated as

$$R_{out} = R_C \parallel r_o \cdot (1 + g_m R_E) \quad (3.19)$$

As can be seen in the presence of the emitter resistor, both the intrinsic input and output impedances will increase by a factor  $(1 + g_m R_E)$ , while the gain is reduced by the same factor.

### 3.1.3 Emitter Follower (Common Collector) Topology

Online YouTube lectures:

[Emitter follower, common-based, cascode, active load, maximum gain](#)

As we saw earlier, the base-emitter voltage of the bipolar junction transistor has a weak (logarithmic) dependence on the current and therefore does not

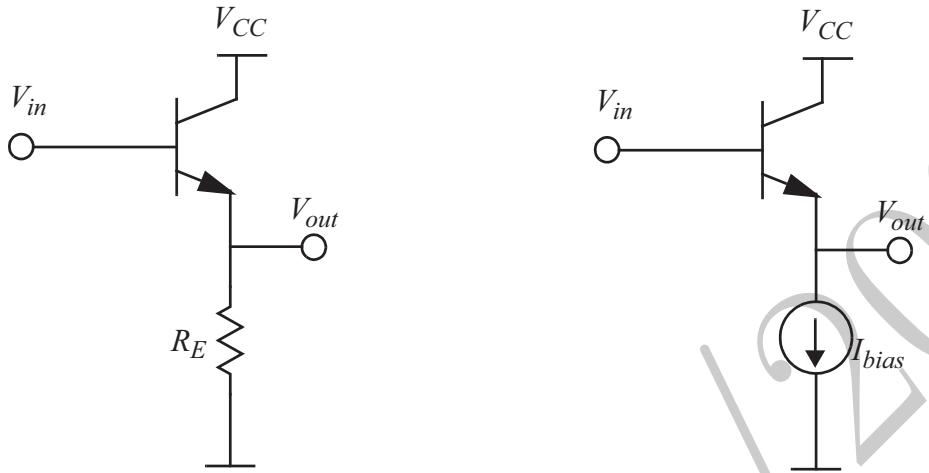


Figure 3.13: The common-collector (emitter follower) stage with a) a resistive load and b) a current source load.

change significantly as the current changes. This property can be used to build a buffer stage that isolates the input from the output, with a level shift equal to  $V_{be}$  of the transistor. Two variations of this stage are shown in Figure 3.13.

In the resistively biased version, the input and the output are related through the following nonlinear relation

$$V_{in} - V_{out} = V_{be} = V_T \ln\left(\frac{\alpha V_{out}}{R_E I_S}\right) \quad (3.20)$$

where  $V_{be}$  does not change a lot and therefore  $V_{out}$  will be a slightly distorted version of  $V_{in}$ . This transfer characteristic can be shown using the following plot of  $V_{out}$  vs.  $V_{in}$  in Figure 3.14.

In the absence of a load resistor, the current source biased version does not introduce distortion as the output and the input voltages are linearly related with a constant level shift, i.e.,

$$V_{in} - V_{out} = V_{BE} = V_T \ln\left(\frac{\alpha I_{bias}}{I_S}\right) \quad (3.21)$$

Now let us look at the small signal model for the transistor and calculate the gain, input, and output impedances of the circuit. Again we use the T-model as it will show the results in a more natural manner. This is illustrated in Figure 3.15, where  $R_B$  is the total series resistance with the base due to the voltage source internal impedance and the base intrinsic series resistance. The effect of this series resistance can be taken into account by noting that the base current

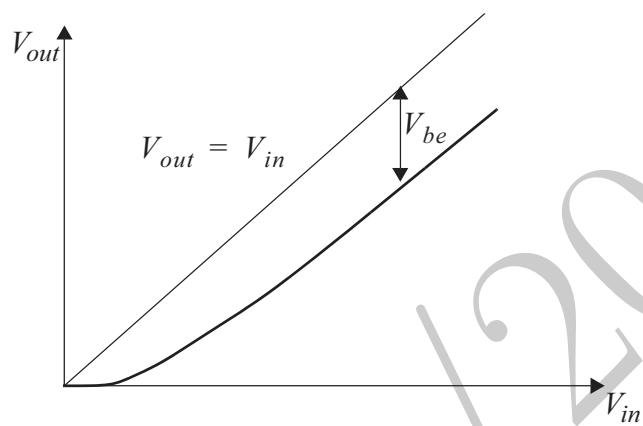


Figure 3.14: The large-signal transfer characteristic of the common-collector (emitter follower) stage.

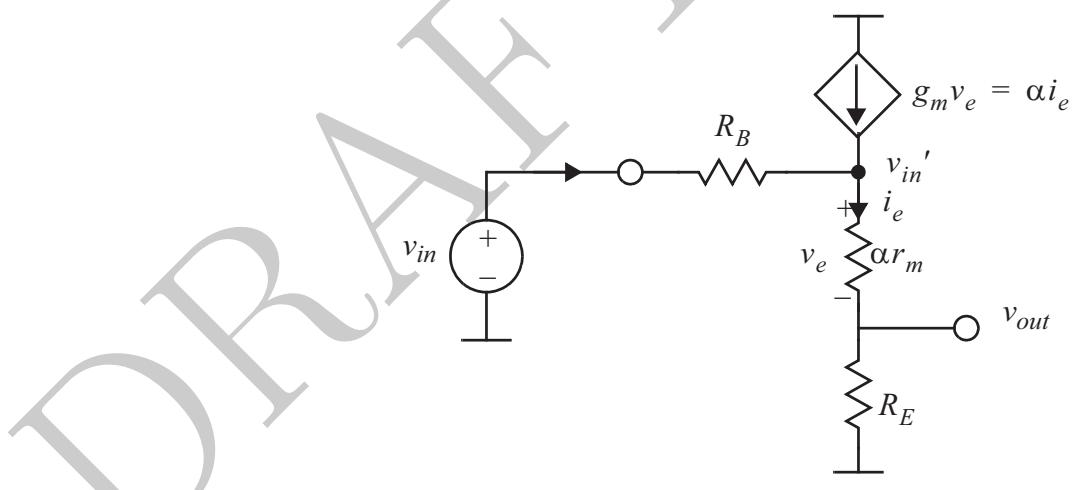


Figure 3.15: The small-signal model for the common-collector (emitter follower) stage.

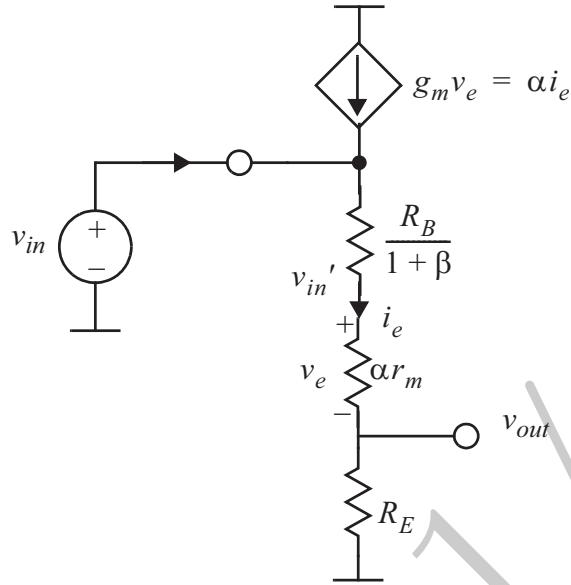


Figure 3.16: The reflection of  $R_B$  to the emitter.

flows through it, and therefore,

$$v_{in} - v'_{in} = R_B i_b = (1 - \alpha) R_B i_e = \frac{R_B}{1 + \beta} i_e \quad (3.22)$$

Based on which, the base resistor can be moved to the emitter branch, given it is divided by  $(1 + \beta)$ . So in general, a resistor can be moved from the base to the emitter and divided by  $(1 + \beta)$  or moved from the emitter to the base and multiplied by the same factor. This is known as reflection rule and can be used to facilitate calculations of gain, input and output impedance, when there are resistors in the base and/or the emitter.

The modified T-model for the emitter follower (common-collector) looks like Figure 3.16.

The gain is simply given by the voltage divider ratio, i.e.,

$$A_V \equiv \frac{v_{out}}{v_{in}} = \frac{R_E}{R_E + \alpha r_m + \frac{R_B}{1 + \beta}} \quad (3.23)$$

which is close to unity as long as  $R_E \gg \alpha r_m$  and  $R_E \gg R_B / (1 + \beta)$ . This allows the stage to be used as a buffer with a voltage gain close to unity. In the case of current source biasing,  $R_E$  is very large and the gain will be one, in agreement with the large signal analysis. One may ask about the usefulness of such a buffer and wonder what the difference between this buffer and a piece of wire may be. To find out, let us look at the input and output impedances of this emitter follower stage.

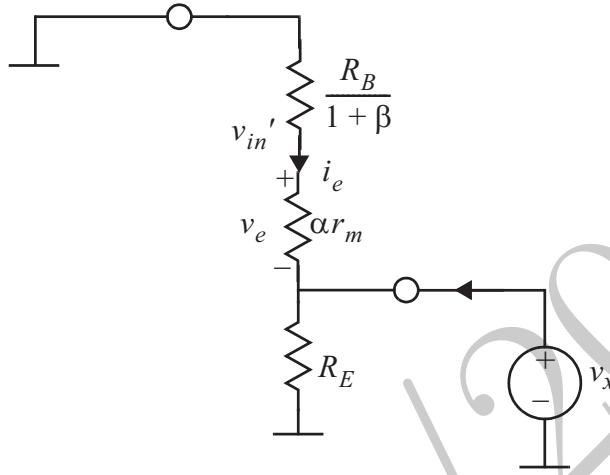


Figure 3.17: The output impedance of the common-collector stage.

The input impedance can be easily calculated using the reflection rule, multiplying all the resistance in the emitter by  $1 + \beta$ , i.e.,

$$R_{in} = (1 + \beta) \cdot \left( \frac{R_B}{1 + \beta} + \alpha r_m + R_E \right) = R_B + r_\pi + (1 + \beta)R_E \quad (3.24)$$

which is essentially the same as the input impedance of the common emitter stage with emitter degeneration as the input circuit looks identical to the common emitter stage.

The output impedance is calculated by setting the input to zero (shorting  $v_{in}$ ) and calculating the output resistance. This is an easy task as in the equivalent model shown in Figure 3.17, the base will be grounded and the output impedance will be given by the parallel combination of  $R_E$  and  $R_B/(1+\beta)+\alpha r_m$ :

$$R_{out} = R_E \parallel \left( \frac{R_B}{1 + \beta} + \alpha r_m \right) \quad (3.25)$$

The calculated input and output resistance of the emitter-follower circuit explain the reason to use it as a buffer. It demonstrates a high input resistance (the same as common emitter) and a very small output resistance as  $r_m$  is usually small for typical values of  $I_C$ . These are the characteristics of a good voltage buffer.

### 3.1.4 Common-Base Topology

Another way of using a bipolar junction transistor is to use it as a common base stage. In this topology, base is biased at a constant voltage to ensure operation in the forward active region and the input is applied to the emitter, as in Figure 3.18.

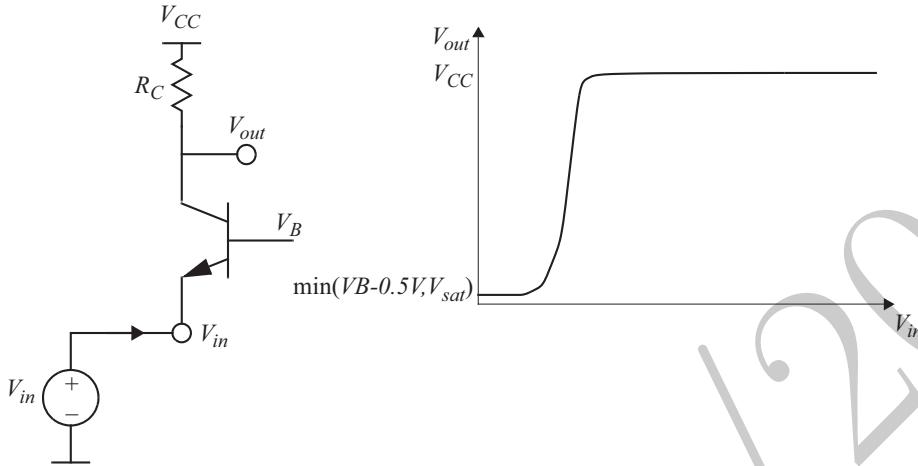


Figure 3.18: The common-base stage.

As can be seen from a simple dc analysis the stage is non-inverting and the output has the same polarity as the input. The gain of the stage can be easily calculated by using the T equivalent model for the transistor and applying an input of  $v_{in}$  to the base, as depicted in Figure 3.19.

The emitter current is simply given by

$$i_{e_0} = -\frac{v_{in}}{\alpha r_m} \quad (3.26)$$

ignoring the output resistance, the output voltage is given by

$$v_{out} = -\alpha i_e R_C = g_m R_C \quad (3.27)$$

which has almost the same absolute value as the gain of the common emitter stage. This should not be surprising as we are applying the input to the base-emitter junction and take the output from the collector. Also, the sign can be explained by noting that the input voltage is  $v_{eb}$  instead of  $v_{be}$  in the common-emitter amplifier.

The input impedance of the common-base stage is  $\alpha r_m$  as can be seen from the small signal model. The common base stage shows a very small input impedance and therefore is useful when the drive is a current sources as the input current will be divided between the parallel Norton resistor of the source and the input impedance of the stage and a smaller input resistance will increase the fraction of the input current entering the amplifier, as shown in Figure 3.20.

The equivalent circuit for the output resistance of the common base stage is essentially identical to the output impedance of the common emitter stage

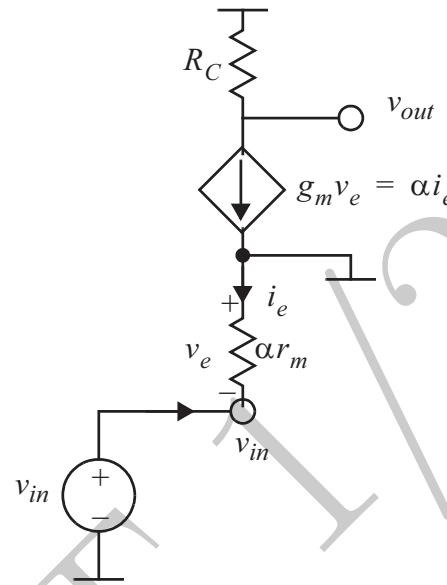


Figure 3.19: The small-signal equivalent for the common-base stage.

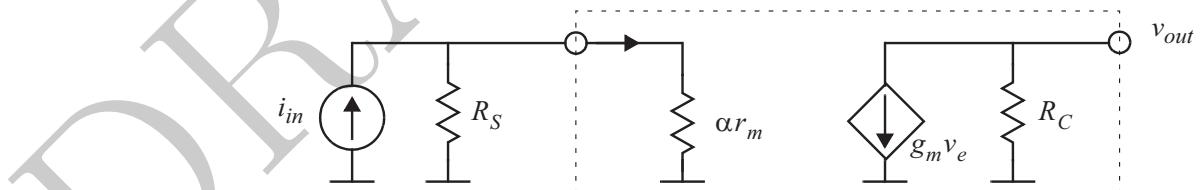


Figure 3.20: The small-signal equivalent for the common-base stage with a source resistance,  $R_S$ .

with emitter degeneration, with the source resistance replacing the emitter resistance, and therefore depends on the source resistance. The output resistance is therefore given by (3.18), i.e.,

$$R_{out} = R_C \parallel r_o \cdot \left( \frac{1 + g_m R_S}{1 + g_m R_S / \beta} \right) \quad (3.28)$$

As can be seen for small source resistance, it will reduce to  $R_C \parallel r_o$ , while for a current source drive (infinite  $R_S$ ) the output resistance is given by  $R_C \parallel \beta r_o$ .

As can be seen the maximum intrinsic gain of a common-base driven by a current source is

$$|A_{max}| = \beta g_m r_o = \beta \frac{V_A}{V_T} = \frac{\beta}{\eta} \quad (3.29)$$

which is  $\beta$  times larger than the maximum intrinsic gain of a common emitter stage. To be able to exploit this interesting characteristic of a common base stage, we need to drive it with a stage whose output can be well approximated as a current source and its input has a fairly high impedance. Both of these requirements are satisfied by a common-emitter stage. The combination of a common emitter and a common base stage results in a cascode stage.

### 3.1.5 Cascode Stage

Online YouTube lectures:

[Two transistor stages: cascade, folded cascode, active load, Darlington, current mirror](#)

A common emitter stage driving the emitter of a common base stage is called a cascode. This stage is shown in Figure 3.21. From small signal point of view, this stage looks like Figure 3.22.

The gain can be calculated by noting that the collector current of  $Q_2$  is  $\alpha$  times its emitter current and the gain is therefore given by

$$A_V \equiv \frac{v_{out}}{v_{in}} = -\alpha g_m (R_C \parallel \beta r_{o2}) \quad (3.30)$$

and the input and output resistances are simply given by

$$R_{in} = r_\pi \quad (3.31)$$

$$R_{out} = R_C \parallel \beta r_{o2} \quad (3.32)$$

Therefore the maximum gain will be given by  $\beta/\eta$ .

There is another reason for using a common base or cascode stage and it is their superior frequency response due to the absence of a capacitance between the input and the output terminal of the common base stage. This issue will be discussed in more details later on.

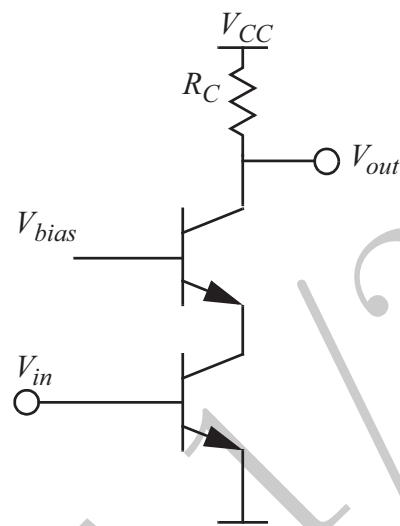


Figure 3.21: The basic BJT cascode stage.

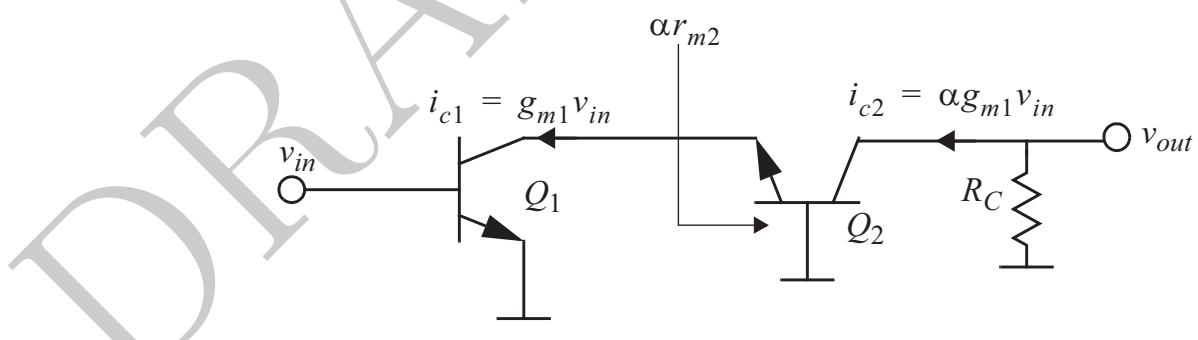


Figure 3.22: The voltage gain calculation of the basic BJT cascode stage.

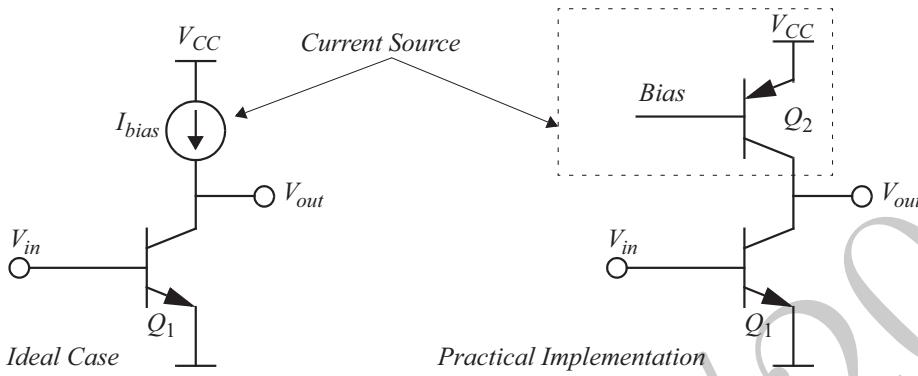


Figure 3.23: The common-emitter stage with an active load.

### 3.1.6 Active Load

As we saw earlier, to obtain a gain close to the maximum achievable gain we need a large ac load resistance. Simply using a very large load resistor will not work as it will force the transistor to enter saturation for very small values of  $I_C$  and hence cause problems with the biasing. A better alternative is a current source, as shown in Figure 3.23.

The current source can be practically realized by biasing the base of a BJT at a well defined voltage so its collector current has a well-defined value. If there were no Early effect, the output current would have been completely independent of the collector voltage. However, in the presence of Early effect, the output current will change with the output voltage and therefore there is a finite resistor in parallel with the current source. This is simply the output resistance of the BJT,  $r_o$ . So Early effect can be taken into account by considering  $r_o$  of the current source in our calculations.

So let us start with gain of the simple emitter follower with a current source load (active load), shown in Figure 3.24.

The effective  $R_C$  is  $r_{op}$  in this case, and therefore the gain is simply given by

$$A_v = -g_m(r_{on} \parallel r_{op}) \quad (3.33)$$

Noting that  $g_m = I_C/V_T$  and  $r_o = V_A/I_C$ , the gain can be written as

$$A_v = -\frac{1}{V_T} \cdot \frac{V_{AN}V_{AP}}{V_{AN} + V_{AP}} = -\frac{1}{\eta_N + \eta_P} \quad (3.34)$$

where  $V_{AN}$  and  $V_{AP}$  are the Early voltages of the NPN and PNP transistors, respectively. It is noteworthy that this gain is independent of the dc current.

As can be seen, using an active load allows us to obtain a gain closer to the maximum intrinsic gain of the device. The gain can be further improved by using a stage with a larger output impedance, e.g., cascode, similar to Figure

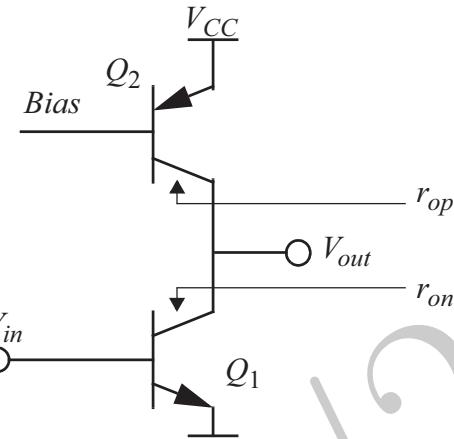


Figure 3.24: The common-emitter stage with a PNP active load.

**3.25.** Replacing the NPN transistor with the cascode combination will increase the gain.

The voltage gain is given by

$$A_v = -g_m(\beta_n r_{on} \parallel r_{op}) \quad (3.35)$$

Now the gain is limited by  $r_{op}$ , and therefore we have to increase it somehow. This can be done using a PNP cascode instead of a single PNP transistor as shown in Figure 3.26.

The voltage gain is given by

$$A_v = -g_{mn}(\beta_n r_{on} \parallel \beta_p r_{op}) \quad (3.36)$$

for the cascode topology shown in the above figure.

## 3.2 MOS Transistor Amplifiers

Online YouTube lectures:

[MOS amplifier stages: Source degeneration, input and output impedances](#)

The  $\pi$ -model for the MOS transistor can also be converted into an equivalent T-model. This time, however, we need to worry about the body effect and the extra dependent current source due to back-gate effect. As mentioned earlier, the MOS transistor can be thought of as the parallel combination of two transistors, a MOSFET formed by the channel, oxide and the gate electrode and a JFET between the bulk and the channel controlled by the bulk-source voltage,  $V_{BS}$ . Figure 3.27 shows the steps involved in this conversion.

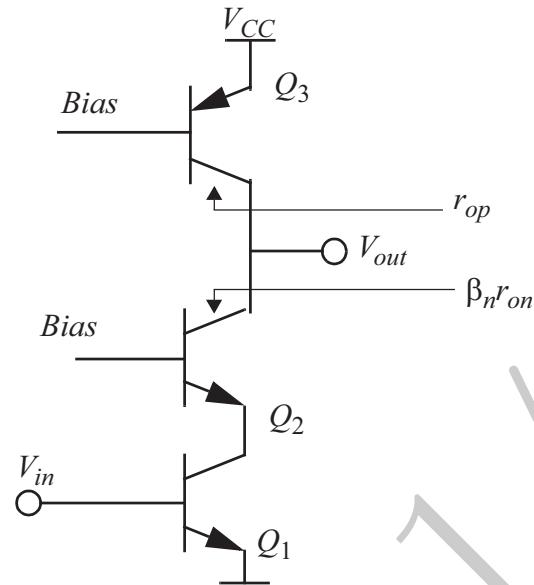


Figure 3.25: The cascode stage with a PNP active load.

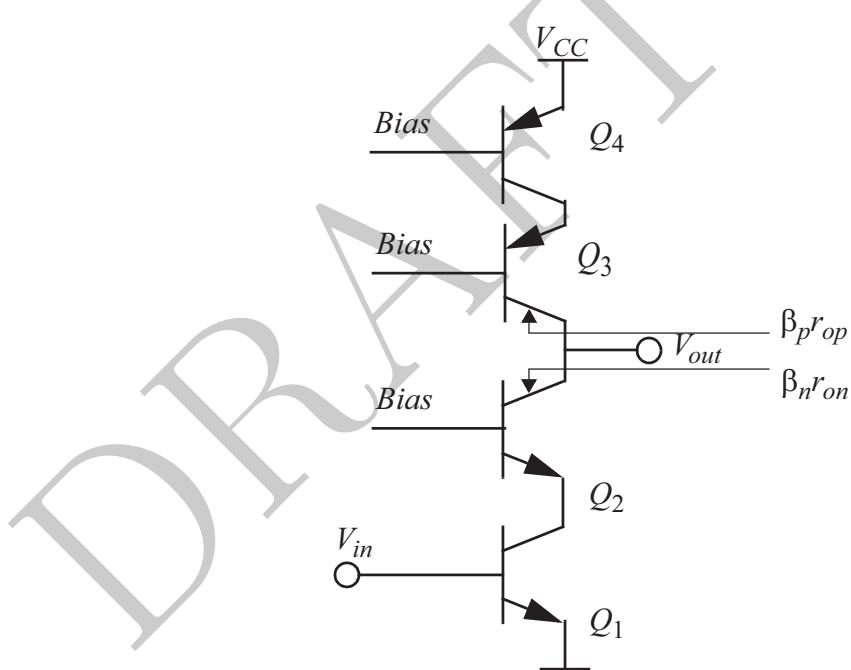


Figure 3.26: The cascode stage with a PNP cascode active load.

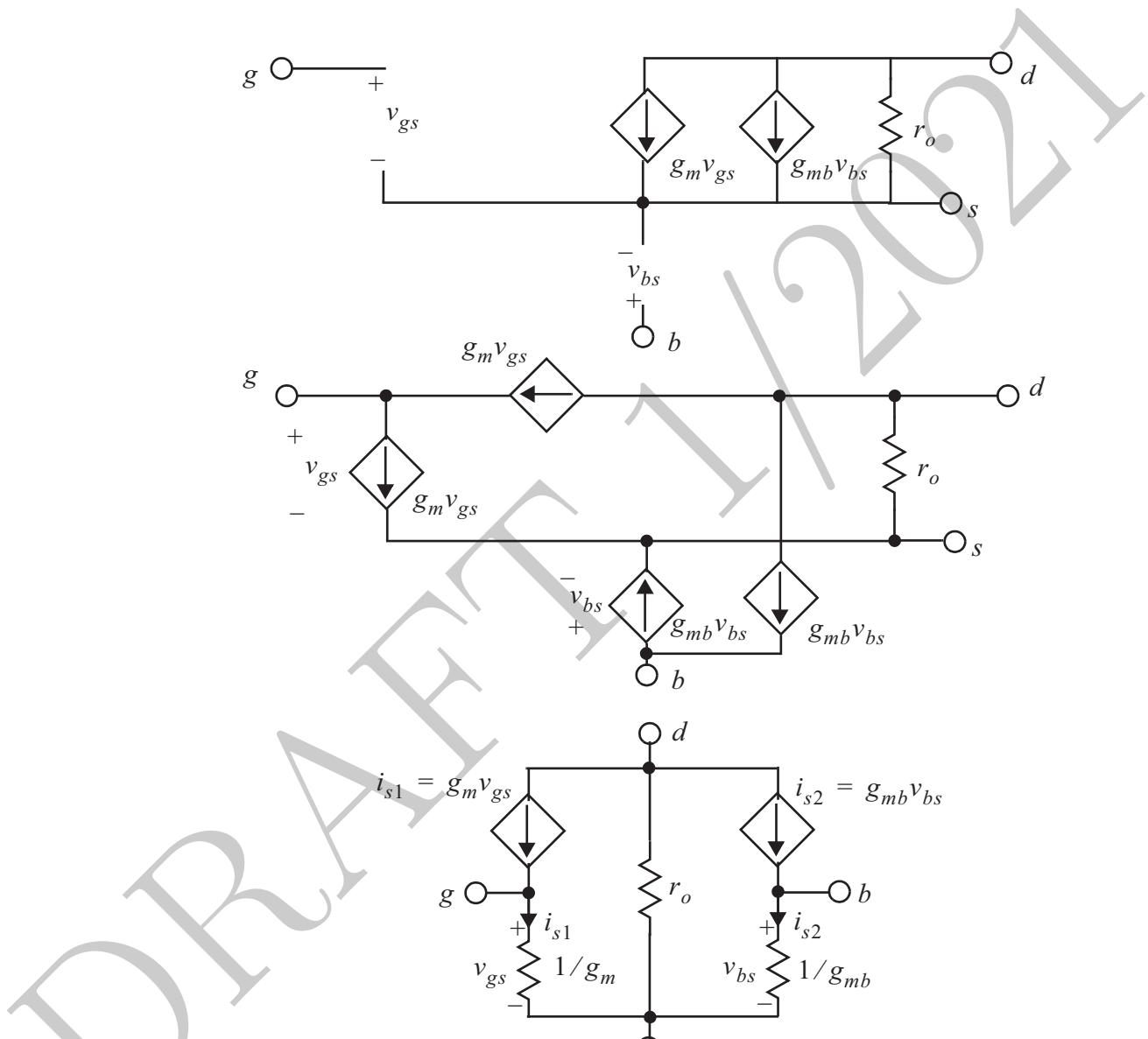
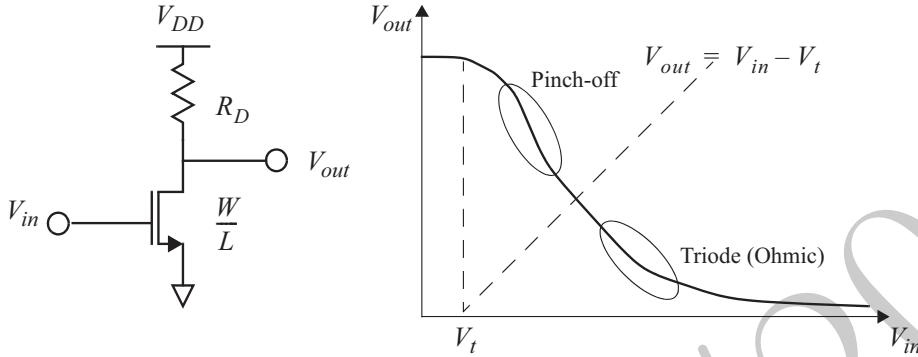


Figure 3.27: Conversion of the  $\pi$ -model to T-Model for a MOSFET.


 Figure 3.28: The basic common-source stage with its  $V_{in}$ - $v_{out}$  characteristics.

As can be seen, there are two T-branches which signifies that there are effectively two transistors in parallel. We will use this equivalent model extensively to analyze MOS transistor circuits.

### 3.2.1 Common Source Topology

Common source is the MOS counterpart of common emitter amplifier. Unlike bipolar processes, in most practical MOS process technologies we don't have access to well controlled resistors. Hence, although we will show some conceptual circuits with resistors, the ultimate circuits will be built mostly with different types of transistor loads.

Let's start with a simple common source amplifier shown in Figure 3.28. Assuming for the time being that the transistor operates in the long channel mode of operation, the large signal transfer characteristic can be calculated. By taking the different regions of operation into account. Ignoring the subthreshold current, for  $V_{in}$  smaller than  $V_t$  the transistor is effectively off. Above  $V_t$  the transistor will start conducting some current. It will start in the pinch-off region, and the output voltage will be related to the input through

$$V_{out} = V_{DD} - R_D I_D = V_{DD} - R_D \frac{\mu_n C_{ox}}{2} \frac{W}{L} (V_{in} - V_t)^2 \quad (3.37)$$

As you can see the input-output have a quadratic dependence. As the input is raised, at some point the pinch-off condition fails to hold and  $V_{in} - V_t > V_{out}$ . At this point the transistor will start operating in the triode region and the output and the input will be related through

$$V_{out} = V_{DD} - R_D I_D = V_{DD} - R_D \mu_n C_{ox} \frac{W}{L} [(V_{in} - V_t)V_{out} - V_{out}^2] \quad (3.38)$$

which can be solved for  $V_{out}$  in terms of  $V_{in}$ . This can also be seen in the

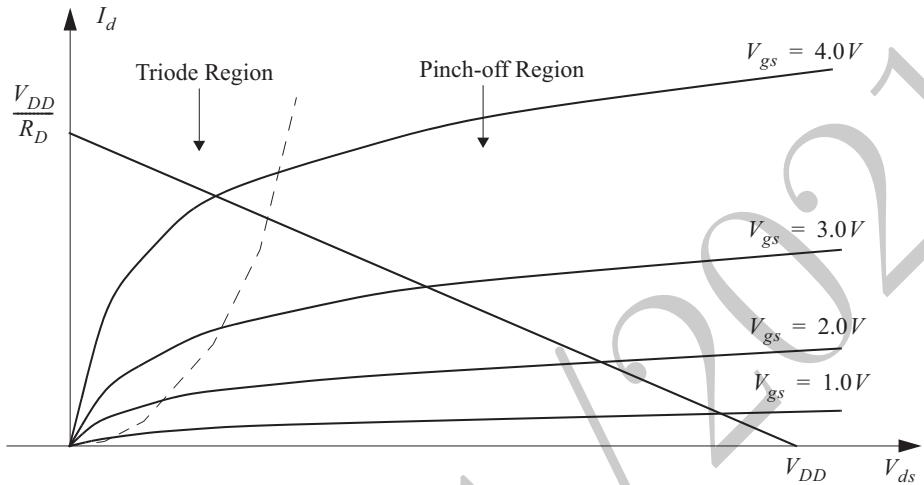


Figure 3.29: The load lines for the basic common-source stage of Figure 3.28.

context of load lines in a manner similar to the bipolar transistor, as illustrated in Figure 3.29.

In the load line picture shown above, the intersection between the load line and device  $I$ - $V$  curves determines the operation point and hence the output voltage. As the input voltage is increased, we move to the  $I$ - $V$  curves for larger  $V_{gs}$  and therefore the operation point moves to the left. This means a reduction in the output voltage given by (3.37) and (3.38).

The gain of the amplifier can be found using the large signal transfer characteristic of (3.37) by calculating the derivative of the output voltage with respect to the input, i.e.,

$$A_v \equiv \frac{\partial V_{out}}{\partial V_{in}} = -R_D \mu_n C_{ox} \frac{W}{L} (V_{in} - V_t) \quad (3.39)$$

Noting that  $g_m = \mu_n C_{ox} W / L (V_{in} - V_t)$ , the expression for gain can be rewritten as

$$A_v = -g_m R_D \quad (3.40)$$

A similar result can be obtain by using the small signal model for this amplifier, as depicted in Figure 3.30.

Note that the back-gate branch is not shown since both source and bulk are grounded and therefore  $v_{bs}$  and  $i_{s2}$  are both zero. As can be seen based on the small signal model the gain of the amplifier is given by:

$$A_v = -g_m (r_o \parallel R_D) \quad (3.41)$$

which reduces to (3.40) if we ignore channel length modulation.

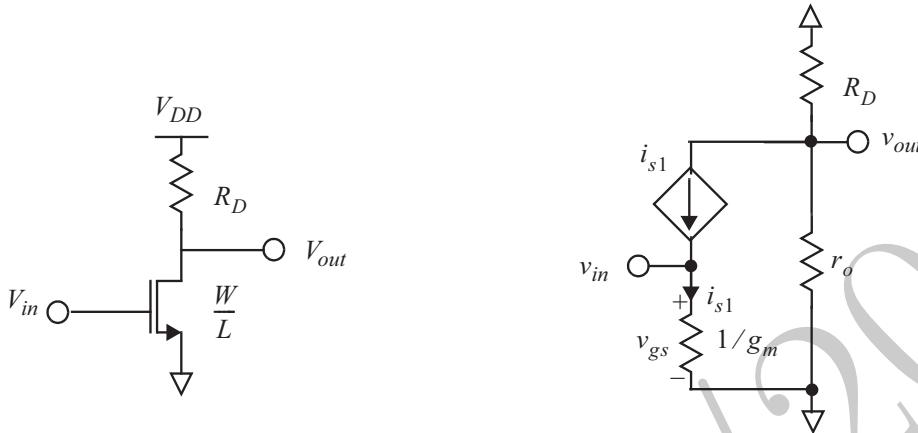


Figure 3.30: The small-signal model of the basic common-source stage of Figure 3.28.

As we mentioned earlier, we usually try to avoid using resistors in a MOS process technology as their values are not usually well controlled. However, let's consider the following common-source amplifier with resistive source degeneration. This topology is not as common as emitter degeneration due to the smaller intrinsic gain of MOS transistors, but it may be used in some applications. Its importance lies mostly in its predictive power when a more complicated circuit can be reduced to this simpler building blocks. A typical common-source amplifier with source degeneration is shown in Figure 3.31.

The amplifier can be analyzed using the small-signal model. Let us first ignore  $r_o$  and calculate the gain. The calculated gain will be valid as long as  $R_D \ll r_o$ . The equivalent small signal model will look like Figure 3.32.

This time back-gate branch cannot be ignored, as although bulk is still at ac ground, source can have a non-zero ac voltage and therefore the back-gate branch should be taken into account. Bulk is grounded and consequently the circuit on the left can be converted to the circuit on the right. The current,  $i_{s1}$  is given by

$$i_{s1} = v_{in} \cdot \frac{g_m(g_{mb} + 1/R_S)}{g_m + g_{mb} + 1/R_S} = v_{in} \cdot \frac{g_m(1 + g_{mb}R_S)}{1 + (g_m + g_{mb})R_S} \quad (3.42)$$

The currents  $i_{s1}$  and  $i_{s2}$  are related by the current divider ratio,

$$i_{s2} = -\frac{R_S}{R_S + 1/g_{mb}} \cdot i_{s1} = -\frac{g_{mb}R_S}{1 + g_{mb}R_S} \cdot i_{s1} \quad (3.43)$$

The output voltage is related to the input by the sum of the total current at the output:

$$v_{out} = -(i_{s1} + i_{s2})R_D \quad (3.44)$$

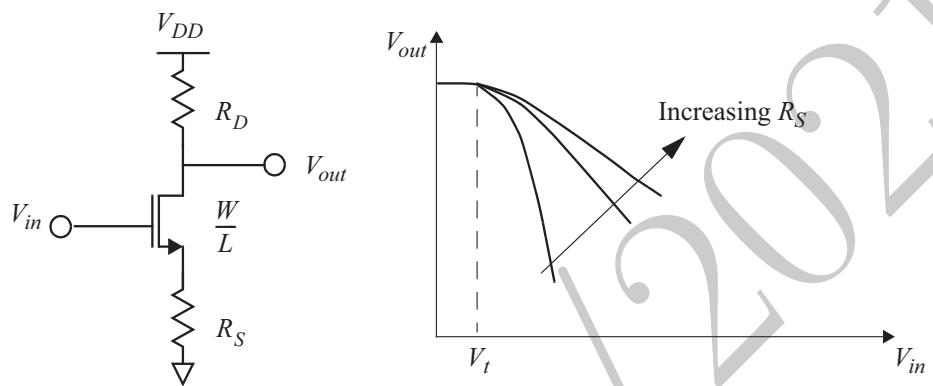


Figure 3.31: The source-degenerated common-source stage with its  $V_{in}$ - $v_{out}$  characteristics.

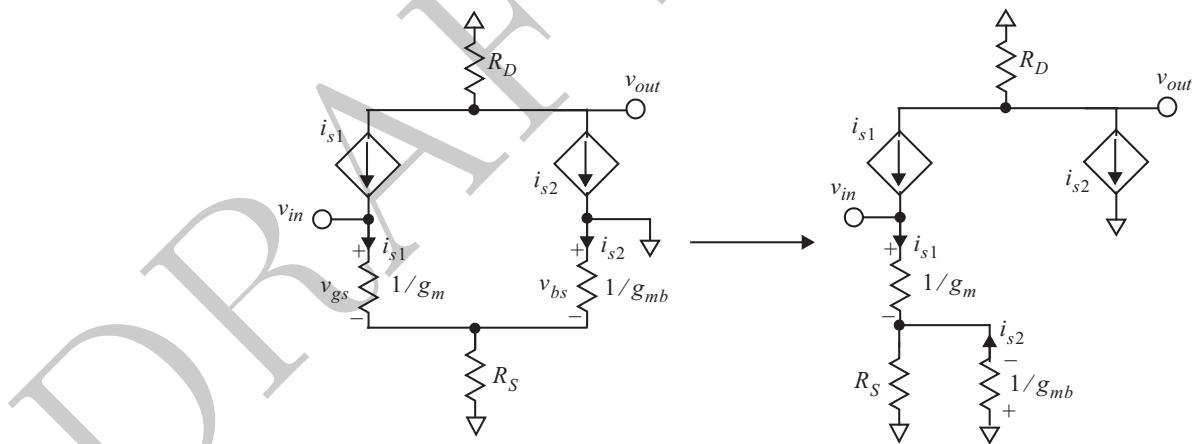


Figure 3.32: The small-signal model of the source-degenerated common-source stage of Figure 3.31.

Combining (3.43) and (3.44), we have

$$v_{out} = -i_{s1}R_D \left(1 - \frac{g_{mb}R_S}{1 + g_{mb}R_S}\right) = -i_{s1}R_D \left(\frac{R_D}{1 + g_{mb}R_S}\right) \quad (3.45)$$

which combined with (3.42) results in the following expression for the gain:

$$A_v \equiv \frac{v_{out}}{v_{in}} = -\frac{g_m R_D}{1 + (g_m + g_{mb})R_S} = -\frac{R_D}{r_m + (1 + \chi)R_S} \quad (3.46)$$

Due to source degeneration, the gain of the degenerate amplifier is lower by a factor of  $1 + (g_m + g_{mb})R_S$  as compared to the non-degenerate case. As can be seen, the gain is more or less given by the ratio of the total drain resistance to the total source resistance with a minor modification. Unlike the bipolar case, there is no  $\alpha$  in this expression which signifies that the low-frequency drain and source currents in a MOS transistor are exactly the same. Also the source resistance is scaled by  $1 + \chi$ , while  $r_m$  and  $R_D$  remain the same.

The input impedance of this stage can be determined by finding the ratio of the input voltage to the input current. However, in the case of MOS amplifier the dependent current source is exactly equal to the current through  $r_m$ . Therefore the small signal low frequency input current is zero. This means that the low frequency input impedance (i.e., input resistance) of this stage is infinity in agreement with the observation that no current flows through the gate.

The output resistance can also be calculated by setting the input source to zero (ac grounding the gate) and finding the ac current through a test voltage source at the output. To simplify our calculation we will ignore  $R_D$  as we know it will be in parallel with the intrinsic output resistance of the stage. This arrangement is shown in Figure 3.33.

The noting that  $r_o$  is much larger than  $g_m$ , the current through  $r_o$  will be approximately given by

$$i_1 \approx \frac{v_x}{r_o} \quad (3.47)$$

Therefore the current through  $1/(g_m + g_{mb})$  is given by

$$i_{s1} + i_{s2} = -i_1 \cdot \frac{g_m + g_{mb}}{g_m + g_{mb} + 1/R_S} = -\frac{v_x}{r_o} \cdot \frac{g_m + g_{mb}}{g_m + g_{mb} + 1/R_S} \quad (3.48)$$

and hence the current through the test voltage source at the output is given by

$$i_x = i_1 + (i_{s1} + i_{s2}) = \frac{v_x}{r_o} \cdot \frac{1}{1 + (g_m + g_{mb})R_S} \quad (3.49)$$

Therefore, the output resistance of this stage is

$$R_{out} = R_D \parallel r_o [1 + (g_m + g_{mb})R_S] = R_D \parallel r_o [1 + (1 + \chi)g_m R_S] \quad (3.50)$$

Due to the source degeneration, the ac gain is reduced by  $1 + (g_m + g_{mb})R_S$  and the output resistance is increased by the same factor.

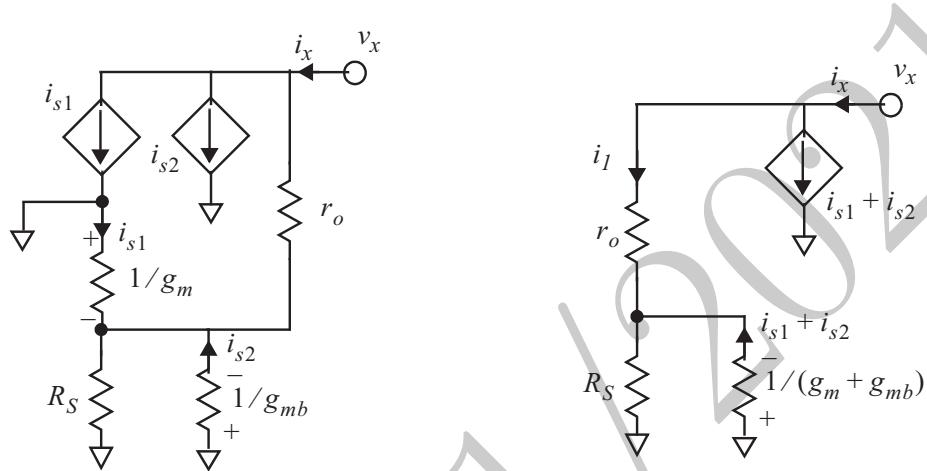


Figure 3.33: The small-signal model of the source-degenerated common-source stage of Figure 3.31 for output impedance calculation.

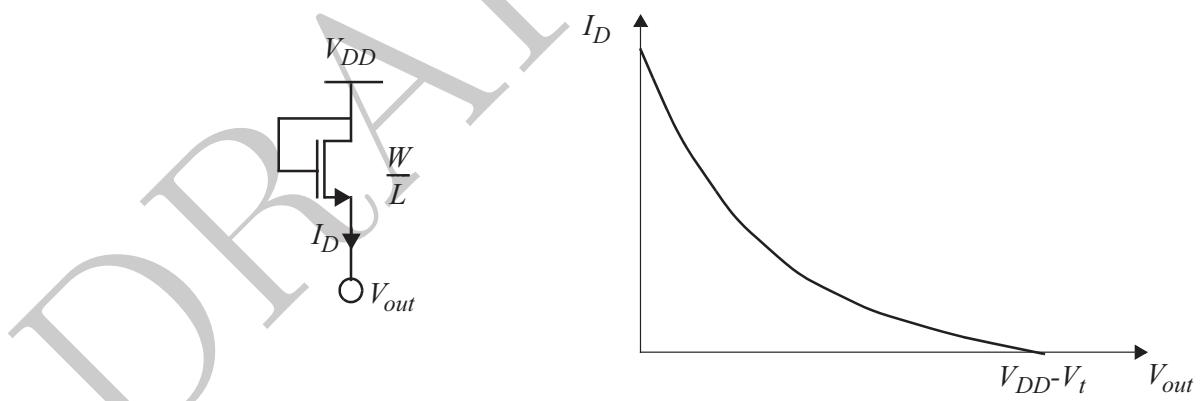


Figure 3.34: The diode connected NMOS as a load.

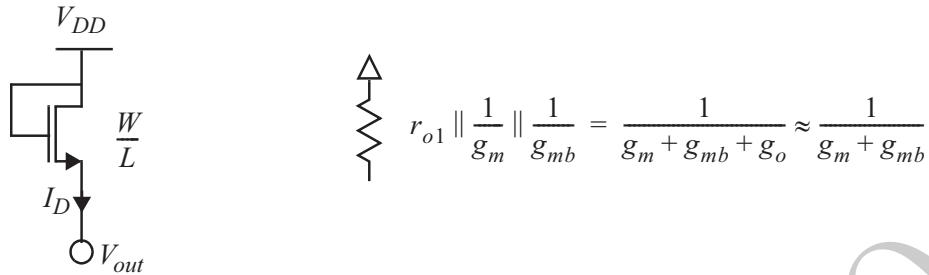


Figure 3.35: The small-signal model of the diode connected NMOS load.

As mentioned earlier, resistors are not controlled very well in standard MOS process technologies and therefore, it is preferable to use transistors as the load. One option is to use a 'diode-connected' NMOS as a load, as shown in Figure 3.34.

Noting that for a positive threshold voltage, the transistor cannot enter triode region because the drain and gate are always at the same potential, the relationship between the source voltage and drain current is given by:

$$I_D = \frac{\mu_n C_{ox}}{2} \frac{W}{L} (V_{DD} - V_{out} - V_t)^2 \quad (3.51)$$

for  $V_{DD} - V_{out} \geq V_t$ , which is depicted in the above load line picture.

The small signal model for this load also contains some very useful information. In the small signal model, the drain, gate and bulk are all grounded, therefore both current sources have a short circuit across and the equivalent circuits looks like Figure 3.35.

This diode-connected NMOS can be used as the load for a common-source NMOS as shown in Figure 3.36.

When the input voltage is increased, the operation point moves from right to left. As can be seen, the gain is not going to be very large as the load line is rather steep. Another way of seeing this is by equating the currents through  $M_1$  and  $M_2$ , i. e.,

$$\sqrt{I_{D1}} = \sqrt{\frac{\mu_n C_{ox}}{2}} \sqrt{\frac{W_1}{L_1}} (V_{in} - V_t) = \sqrt{\frac{\mu_n C_{ox}}{2}} \sqrt{\frac{W_1}{L_1}} (V_{DD} - V_{out} - V_t) = \sqrt{I_{D2}} \quad (3.52)$$

which results in the following large signal transfer function as long as the input transistor remains in pinch-off region,

$$V_{out} = V_{DD} - V_t - \sqrt{\frac{W_1/L_1}{W_2/L_2}} (V_{in} - V_t) \quad (3.53)$$

Note that threshold voltage is different for the transistors due to body effect. Based on the above equation and ignoring body-effect (equal  $V_t$ s), the small

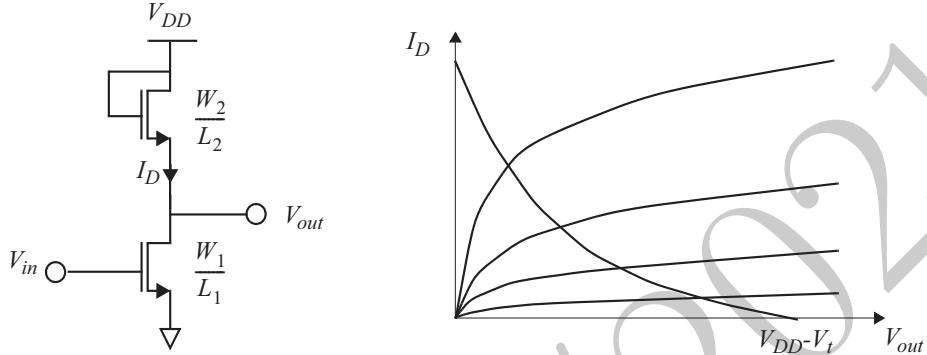


Figure 3.36: A common-source stage with the diode connected NMOS load.

signal gain is simply the square root of the ratio of the  $W/L$ s. Note that the gain is independent of the operation point(e.g.,  $ID$ ).

Now let us use the small signal model and take body effect into account. The equivalent small signal model for this amplifier will look like Figure 3.37.

Noting that  $1/g_m, 1/g_{mb} \ll r_o$ , the small-signal gain can be calculated as:

$$A_v \equiv \frac{v_{out}}{v_{in}} = -\frac{g_{m1}}{g_{m2} + g_{mb2}} = -\frac{g_{m1}}{g_{m2}} \cdot \frac{1}{1 + \chi} = \sqrt{\frac{W_1/L_1}{W_2/L_2}} \cdot \frac{1}{1 + \chi} \quad (3.54)$$

As can be seen, the small signal gain is independent of the current and only depends on the ratio of geometric parameters and the strength ratio of the back gate and the main gate,  $\chi$ . This is a good feature as it allows for a rather constant gain over a large range of input and output voltages. Gain is also independent of process parameters such as  $C_{ox}$  and  $\mu$ .

To maximize the gain we need to have a large ratio between the  $W/L$ s of the input and load transistors. However, even if this ratio is around 100, the gain will be smaller than 10 due to the square root dependence. This is a rather undesirable characteristic as to obtain a large gain the input transistor should be very wide and short as opposed to a thin and long load device. These unbalanced ratios can create difficulties during the layout of the actual circuit.

The question is how to increase the gain without having to use excessively large aspect ratios. To answer this question, we first have to identify the reason for this small gain. Ignoring body-effect, the gain is essentially given by the ratio of the input transistor transconductance to the transconductance of the output transistor. Based on this observation we note that to increase the gain we need to lower the transconductance of the load transistor. In addition to reducing its  $W/L$  ratio, this can be done by lowering its current. However, this should be done without lowering the current through the input transistor since this will reduce  $g_m$ . One solution would be to 'steal' some of the current away

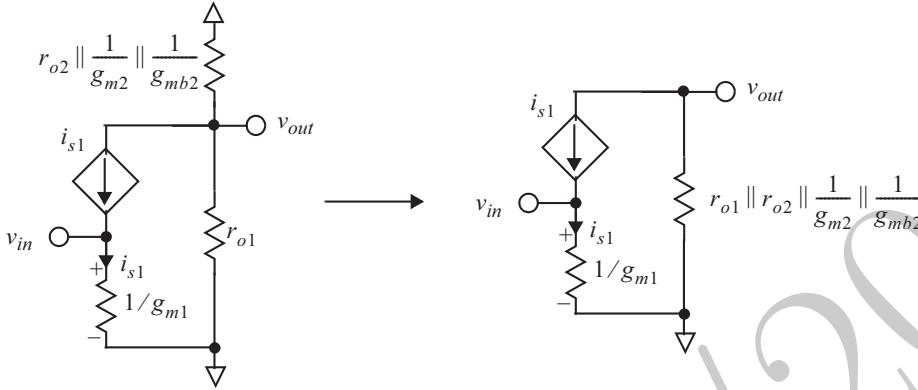


Figure 3.37: The small-signal model of the common-source stage with the diode connected NMOS load.

from the load transistor and thereby decreasing its transconductance. This can be done, by introducing a current source in parallel with the load transistor as shown in Figure 3.38.

It may seem that by increasing the ratio of  $I_{steal}$  current to the load current, the load resistance and hence the gain can be increased without bound. The extreme case is when the load is completely removed and the transistor is loaded with a PMOS current source (active load). An example of such circuit is shown in Figure 3.39.

The gate of the PMOS transistor is biased at a constant voltage and it acts as a current source. As the bulk and the source of both NMOS and PMOS transistors are at a constant potential, there is no body effect involved.

The large signal input-output behavior of the circuit can be understood by tracking the regions of operation transistors go through as the input voltage is increased. For input voltages smaller than  $V_{tN}$ , NMOS transistor is off and PMOS is in deep triode region and no current will flow in the circuit. As the input voltage is raised above  $V_{tN}$ , the NMOS will enter pinch-off region while PMOS is still in triode region. Once enough input voltage is applied, both transistors will be in pinch-off which is the desired mode of operation for amplification. As the input voltage is further increased, at the output voltage will further drop and the NMOS will enter triode region. This behavior can also be understood by using the load lines shown in Figure 3.40.

The small signal model for this circuits looks like Figure 3.41.

It is easy to see from this model that the small signal gain of this stage is given by

$$A_v \equiv \frac{v_{out}}{v_{in}} = -g_{mn}(r_{on} \parallel r_{op}) = -\frac{g_{mn}}{g_{on} + g_{op}} \quad (3.55)$$

The lack of any  $g_{mb}$  in the above expression also indicates that the gain

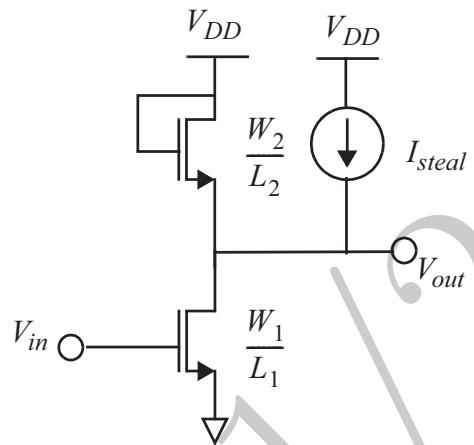


Figure 3.38: The bleed current in parallel with the diode connected NMOS load.

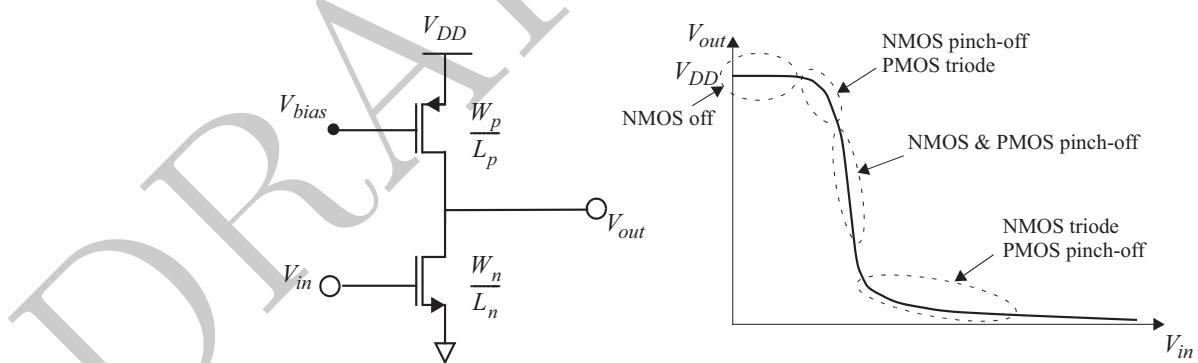


Figure 3.39: A common-source with a PMOS active load.

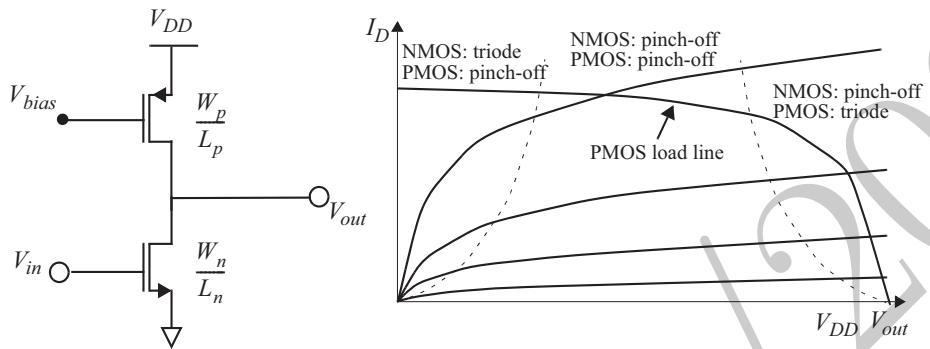


Figure 3.40: The load lines of the common-source with a PMOS active load.

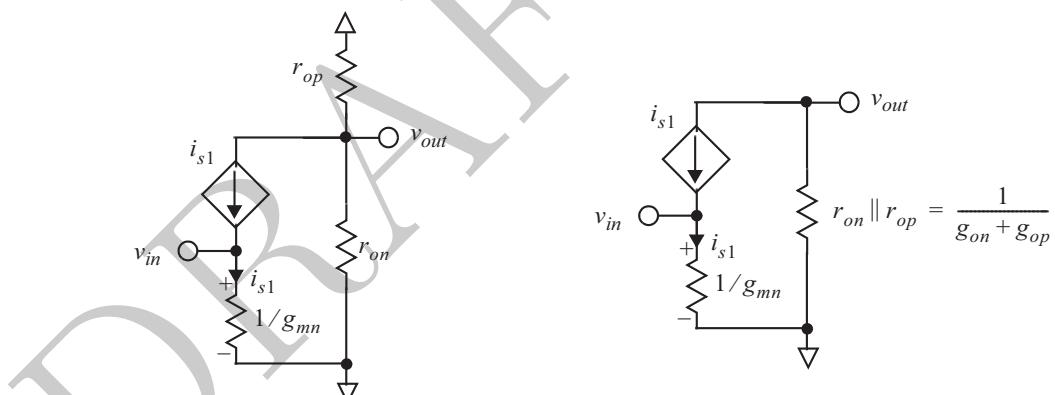


Figure 3.41: The small-signal model of the common-source with a PMOS active load.

is not affected by body effect in agreement with our previous observation. As can be seen from the above expression, the maximum achievable gain from this stage is determined by the  $g_{mr_o}$  of the transistor. Using the expressions for transconductance and output resistance in a long channel MOSFET, the gain can be expressed as:

$$|A_{v,max}| = g_m r_o = \sqrt{2\mu_n C_{ox} \frac{W}{L} I_D} \cdot \frac{L}{I_D} \left( \frac{dx_x}{dV_{DS}} \right)^{-1} \quad (3.56)$$

The small signal gain can be increased by increasing the W and L of the transistor, i.e., increasing its gate area. As we will see later, this is in direct conflict with faster operation of the transistor as larger gate area means larger gate capacitance. Equation (3.56) also indicates that the gain is inversely proportional to the square root of the drain current. This may seem counter-intuitive at first but the dependence of the output resistance on drain current is stronger than the dependence of the transconductance on it. Although this equation may suggest that the gain approaches infinity for zero drain current, at very low current level the transistor will be in sub-threshold region and hence its transconductance will behave more like a bipolar transistor and will be proportional to  $I_D$  as opposed to its square root and hence the maximum gain will reach a constant value and will not go to infinity.

The above equation also indicates that the gain is inversely proportional to the width of the depletion region,  $x_d$ , and since  $x_d$  is inversely proportional to the square root of the doping concentration, it will increase with higher doping densities.

### 3.2.2 Source Follower (Common Drain)

Online YouTube lectures:

[\*\*Common-drain, common-gate, cascade, increasing the gain\*\*](#)

Another stage that is sometimes used is source follower which is the MOS counterpart of the emitter follower (common collector) stage. Two different implementations of this stage are shown in Figure 3.42.

Let us start with the resistively loaded version. The equivalent small signal model for this stage is like Figure 3.43.

As can be easily seen from this picture, the small signal gain is given by

$$A_v \equiv \frac{v_{out}}{v_{in}} = \frac{\frac{R_S}{1+g_m R_S}}{\frac{R_S}{1+g_{mb} R_S} + \frac{1}{g_m}} = \frac{g_m R_S}{1 + (1 + \chi) g_m R_S} = \frac{R_S}{r_m + (1 + \chi) R_S} \quad (3.57)$$

which reduces to

$$A_v \equiv \frac{v_{out}}{v_{in}} = \frac{1}{1 + \chi} \quad (3.58)$$

for an ideal current source drive (infinite  $R_S$ ). As can be seen body effect has a rather undesired effect on the gain as there is a limit to how close the gain can get to the ideal gain of unity.

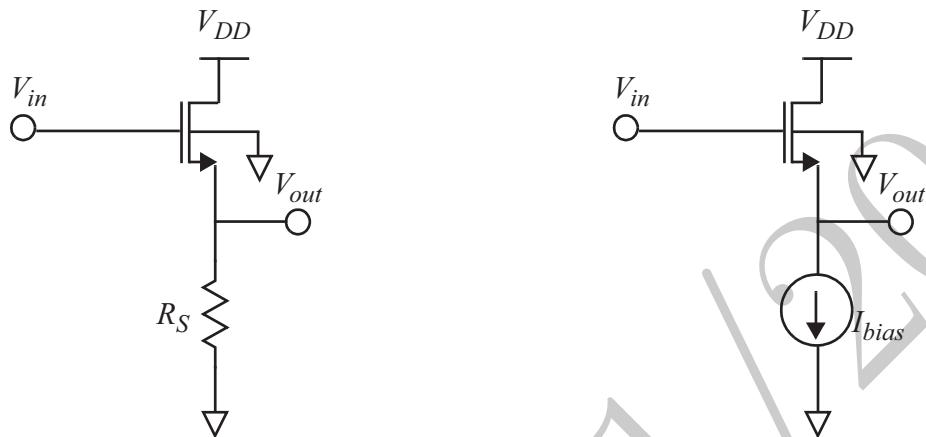


Figure 3.42: The common-drain (source-follower) stage.

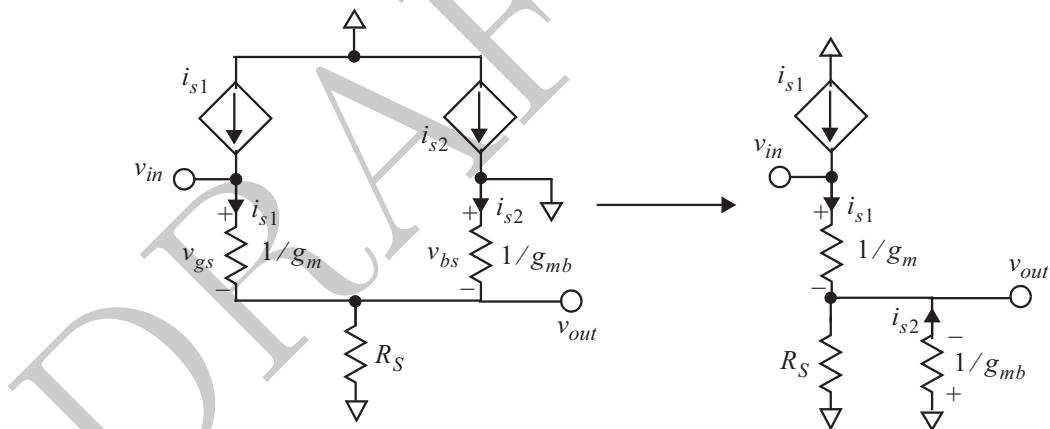


Figure 3.43: The small-signal model of the common-drain (source-follower) stage.

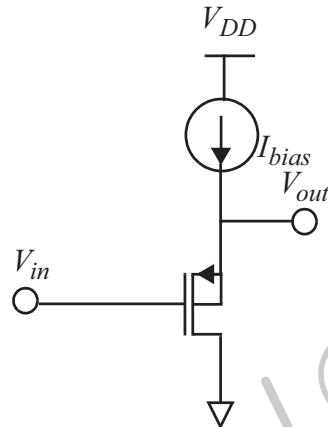


Figure 3.44: A PMOS common-drain (source-follower) stage.

The input impedance is the same as that of the common source and is infinite at low frequencies. The output impedance can easily be calculated from the above equivalent circuit by setting the input terminal to zero. Doing this the output impedance is simply given by

$$R_{out} = R_S \parallel \frac{1}{g_m + g_{mb}} \quad (3.59)$$

As can be seen, it is small for typical values of  $g_m$  and  $g_{mb}$ . The effect of body effect can be eliminated if a PMOS is used instead as the well contact can be connected to the source, maintaining VBS at zero for all values of the input and output voltage as shown in Figure 3.44. Since the dependence of the drain current on the gate-source voltage is not as strong as the dependence of the collector current on the base-emitter voltage in a bipolar transistor, by adjusting the  $W$  and  $L$  of the device in conjunction with the bias current,  $V_{gs}$  can be set to a desired value and the stage can be used as a dc level shifter.

### 3.2.3 Common Gate

The common gate stage is the MOS version of the common base stage. The signal is applied to the source and the output is taken from the drain, as shown in Figure 3.45.

The large signal transfer characteristic can be obtained by writing the expression for the output voltage in terms of the drain current:

$$V_{out} = V_{DD} - R_D I_D = V_{DD} - R_D \frac{\mu_n C_{ox}}{2} \frac{W}{L} (V_{bias} - V_t - V_{in})^2 \quad (3.60)$$

The small signal gain of this stage can be obtained by using the small signal model for the amplifier illustrated in Figure 3.46.

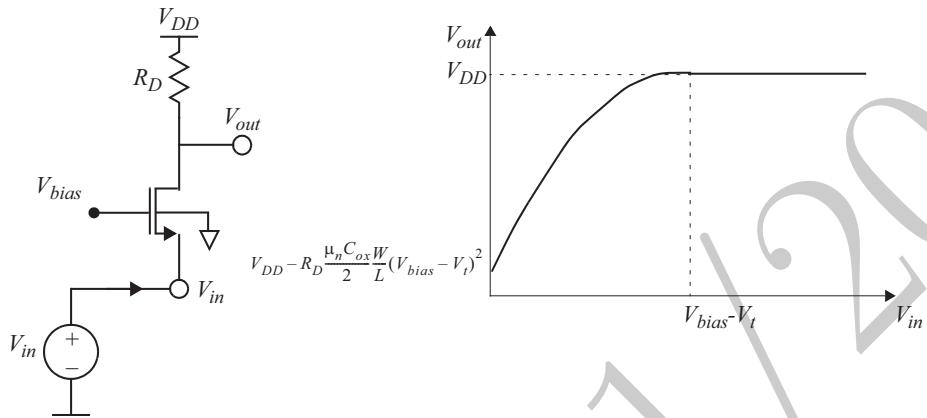


Figure 3.45: The common-gate stage.

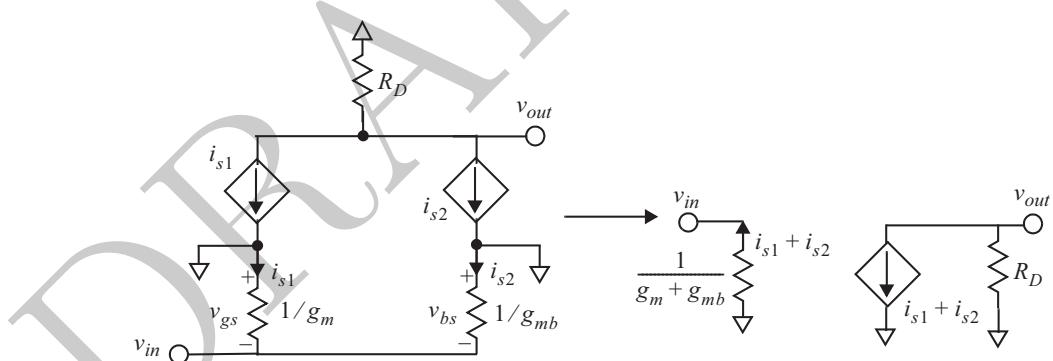


Figure 3.46: The small-signal model for the common-gate stage.

The gain can be calculated by inspection to be

$$A_v \equiv \frac{v_{out}}{v_{in}} = (g_m + g_{mb})R_D \quad (3.61)$$

As can be seen from this expression, in the case of common gate topology, the body effect actually increases the gain and is helpful. This makes intuitive sense because both the main and body transistors are in common gate configuration and share the same load. Therefore, the total transconductance is the sum of the two transconductances.

The output resistance of the common gate amplifier can also be calculated by using the small signal model. For the output resistance, we will consider the more general case of input source with a source resistance  $R_S$ . The small-signal model for the common-gate amplifier is going to be identical to the one used to calculate the output resistance of a common-source amplifier with source degeneration. Therefore the output impedance will be given by (3.50).

The subjects of differential signaling, differential amplifiers in various technologies and their similarities and differences are covered in the following online YouTube lectures:

[\*\*Differential Signaling and differential pair: core concept and large signal behavior \(BJT\)\*\*](#)

[\*\*Differential Amplifiers: MOS, BJT, and ATD \(p.1\)\*\*](#)

[\*\*Differential Amplifiers: MOS, BJT, and ATD \(p.2\)\*\*](#)

[\*\*Differential amplifier with active load. Differential-to-single-ended conversion.\*\*](#)

[\*\*MOS Differential-to-Single-Ended Conversion. Gain Enhancement.\*\*](#)

Some of the biasing concepts are covered in the following online YouTube lectures:

[\*\*Basic Biasing Concepts\*\*](#)

[\*\*Integrated circuit biasing, current mirrors, headroom\*\*](#)

[\*\*Process, Supply, and Temperature Independent Biasing\*\*](#)

[\*\*Scaled bandgap reference, adjustable voltage PVT independent references.\*\*](#)

A brief overview of driver stages is presented in the following YouTube lecture:

[\*\*Driver stages, output stages, Class A and Class B\*\*](#)

A couple of examples of op-amp design and the thought process behind it can be found in the following YouTube lectures:

[\*\*Op-Amp Design: Basic MOS Op-Amp\*\*](#)

[\*\*MOS Op-Amp Design Examples\*\*](#)

[\*\*BJT Op-Amp Design Example\*\*](#)

### 3.3 Resistor Calculations for TTC Method

These results are particularly useful in calculating time- and transfer constants in TTC method in the subsequent chapters.

## Figure Place Holder

Figure 3.47: A transistor stage with arbitrary resistors between its terminals and the ac ground.

Consider the general transistor stage with a resistor  $R_1$  between the base (gate) and ac ground,  $R_2$  between collector (drain) and the ac ground, and  $R_3$  between emitter (source) and the ac ground, as shown in Figure 3.47. We are interested in determining the low-frequency resistance seen between any two terminals of the transistor, as shown in the Figure 3.47. We can show (in Problem XXX) that the base-emitter (or gate-source) resistance,  $R_\pi^0$ , is given by

$$R_\pi^0 = r_\pi \parallel \frac{R_1 + R_3}{1 + g_m R_3} \quad (3.62)$$

The base-collector (or gate-drain) resistance,  $R_\mu^0$  is given by:

$$R_\mu^0 = R_{left} + R_{right} + G_m R_{left} R_{right} \quad (3.63)$$

where

$$R_{left} \equiv R_1 \parallel [r_\pi + (1 + \beta)R_3] \quad (3.64a)$$

$$R_{right} \equiv R_2 \quad (3.64b)$$

$$G_m \equiv \frac{1}{r_m + R_3} = \frac{g_m}{1 + g_m R_3} \quad (3.64c)$$

Note that  $R_{left}$  is the resistance seen between the base (gate) and the ac ground which reduces to  $R_1$  for a MOSFET ( $\beta \rightarrow \infty$ ). Resistance  $R_{right}$  is the resistance between the collector (drain) and ac ground, and finally  $G_m$  is the effective transconductance. XXX double check  $R_\mu^0$  XXX.

The resistance seen between the collector and the emitter (drain and source),  $R_\theta^0$ , is given by

$$R_\theta^0 = \frac{R_2 + R_3}{1 + \frac{1+\beta}{\beta+g_m R_1} g_m R_3} \approx \frac{R_2 + R_3}{1 + g_m R_3} \quad (3.65)$$

where the first expression is exact and the approximation disappears when  $\beta \rightarrow \infty$ . These results come handy in the subsequent chapters and may be worth remembering. Note that  $R_\theta^0$  is not the same as the resistance seen between the collector and ground, namely,  $R_{right}$ .

## Chapter 4

# High Frequency Behavior

We discussed the low frequency behavior of amplifiers in Chapter 3, ignoring the effect of the reactive elements (*e.g.*, capacitors and inductors). However, they are present in all electronic circuits as parasitic elements in the physical realization of the circuit components and interconnects. The reactive elements are also used as intentional elements of the design to improve its performance or to produce qualitatively different behavior (*e.g.*, turning an amplifier into an oscillator). Using these elements is also an important part of the creative design process for high-speed and high-frequency circuits.

In this chapter we develop a set of tools of varying degrees of accuracy and ease of use to analyze and understand high-frequency behavior of circuits. The objective of the analysis should be to provide the designer with a clear understanding of the causal relation among different design parameters, topological choices, and the performance. Direct analysis based on KCL and KVL is not generally very helpful in this kind of approach for several reasons. First of all, no partial results can be obtained before the completion of the analysis. The analysis usually involves solving systems of equations in parametric form that is prone to errors. Assuming we make no mistakes and arrive at the correct final result, we need to simplify them to a tractable form to arrive at some useful conclusions.

We introduce and use the design-oriented method of *Time and Transfer Constants* (TTC) to determine the high-frequency behavior of circuits in this chapter. The salient feature of the TTC approach is its modular nature. It is applied in a successive fashion, with an incremental increase in complexity and accuracy of the results. This allows one to arrive at a first order understanding of the problem through a simple straightforward analysis. Subsequent steps introduce additional terms that build on the original results and improve upon them. While if carried to the end TTC produces exactly the same result as the nodal analysis, the process can be stopped once the desired level of accuracy is achieved.

Another important feature of the *Time and Transfer Constants* method is its ability to identify the dominant limiting effects in the circuit because of its

modular nature. It is obvious that the design effort should be concentrated on the dominant, first-order effects first, before focusing on less-dominant (and usually cosmetic) higher-order modifications.

In the subsequent sections we will discuss the analysis tools in an ascending order of complexity and accuracy. We first talk about some of the general properties of high-frequency transfer functions in Section 4.1. Then, we apply the method of time constants to a first-order system with one energy-storing element in Section 4.2. Next, in Sections 4.3 and 4.4, we will generalize this approach to determine the two most important parameters used to estimate the response of an  $N$ th-order system with the *zero-value time constant* methods for the poles *and* the zeros. The exact response of a second-order system is discussed in Section 4.5. Finally, in Section 4.6, we introduce the TTC method which, when carried to the end, produces the exact transfer function of the circuit using time-constants and transfer constants, and subsumes the earlier techniques. Section ??, deals with how the time-constants, poles, and zeros affect the time-domain response of a circuit<sup>1</sup>.

## 4.1 General Properties of Transfer Functions

Online YouTube lectures:

### [High frequency: transfer functions, lower pass and high pass response.](#)

The transfer function of a linear system with lumped elements can be written as<sup>2</sup>:

$$H(s) = \frac{a_0 + a_1 s + a_2 s^2 + \dots + a_m s^m}{1 + b_1 s + b_2 s^2 + \dots + b_n s^n} \quad (4.1)$$

where all  $a_i$  and  $b_j$  coefficients are real and  $s$  represents the complex frequency. In particular,  $a_0$  is the low frequency (dc) transfer function, as shown in Figure 4.1. Coefficients  $b_i$  have units of seconds to the  $i$ th power and coefficients  $a_i$  have the same units as the transfer function itself times seconds to the  $i$ th power<sup>3</sup>.

<sup>1</sup>All the results in this section are derived based on physical arguments, however, it is conceivable that you are interested in knowing the key results. The derivation of an important result is often labeled so you can skip to the “result” label if you wish. Nonetheless, there is some real value in understanding the derivation to understand the underlying assumptions leading to the final outcome and how they can be applied. If you are *really* comfortable with the deductive approach and rather see the final, most general result first, you might want to go straight to Section 4.6 and come back to the earlier Sections later. However, you will miss a lot of intermediate discussions and results. So do this at your own risk!

<sup>2</sup>In the most general case, the leading term in the denominator should be written as  $b_0$ . However, this introduces a degree of ambiguity in the choice of the coefficients. The only time this becomes an issue is a transfer function that goes to infinity at dc. An example of such a transfer function is the input impedance of a capacitor to ground. In such cases, it is easiest to consider the inverse transfer function (e.g., admittance in the case of the capacitor). Alternatively one can remove the leading 1 in the denominator of (4.1) and modify the coefficient calculations. Nonetheless, we maintain (4.1) in its current form as it results in considerable simplification in most cases and will deal with the pathological cases by evaluating the inverse transfer functions.

<sup>3</sup>For example, if the transfer function is an impedance, the units of  $a_i$  would be  $[\Omega \cdot \text{sec}^i]$ .

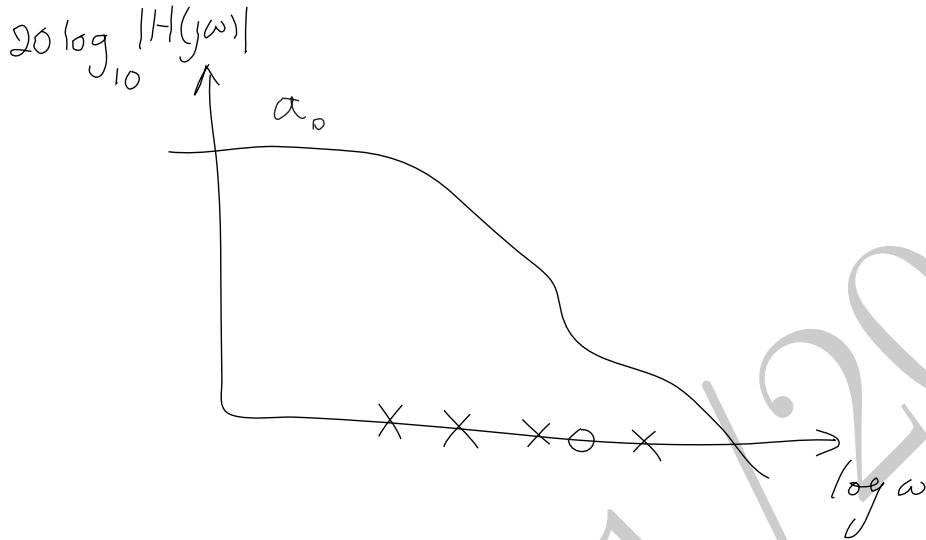


Figure 4.1: A Low Pass Transfer Function.

It is noteworthy that in general the transfer function could refer to the ratio of the voltages and/or currents of any two arbitrary ports of the network, including the ratio of the voltage and current of the same port. In the later case, it is simply the impedance (or admittance) of that port (e.g., input or output impedance).

The Fundamental Theorem of Algebra states that any single-variable polynomial of the degree  $n$  with complex coefficients has exactly  $n$  complex roots. Furthermore, it implies that if all the coefficients are real (as is the case both the numerator and denominator of (4.1)), the roots are either real or appear in complex conjugate pairs<sup>4</sup>. Based on this we can rewrite (4.1) as:

$$H(s) = a_0 \cdot \frac{(1 - \frac{s}{z_1})(1 - \frac{s}{z_2}) \dots (1 - \frac{s}{z_m})}{(1 - \frac{s}{p_1})(1 - \frac{s}{p_2}) \dots (1 - \frac{s}{p_n})} \quad (4.2)$$

where  $z_i$  and  $p_i$  are the pole and zero frequencies and are either real or appear in complex conjugate pairs<sup>5</sup>. The factorization of (4.2) is most suitable to describe

<sup>4</sup>This is equivalent to saying that any single-variable polynomial with *real* coefficients can be factored as a product of first and second order polynomials with real coefficients.

<sup>5</sup>Comparing (4.1) and (4.2) it is apparent that  $a_i$  and  $b_i$  coefficients can be expressed in terms of  $z_i$  and  $p_i$  as follows:

$$\begin{aligned} b_1 &= -\sum_i \frac{1}{p_i}, & \frac{a_1}{a_0} &= -\sum_i \frac{1}{z_i}, \\ b_2 &= \sum_i \sum_{i < j} \frac{1}{p_i p_j}, & \frac{a_2}{a_0} &= \sum_i \sum_{i < j} \frac{1}{z_i z_j} \\ &\vdots & &\vdots \\ b_n &= \frac{(-1)^n}{p_1 p_2 \dots p_n}, & \frac{a_m}{a_0} &= \frac{(-1)^m}{p_1 p_2 \dots z_m} \end{aligned}$$

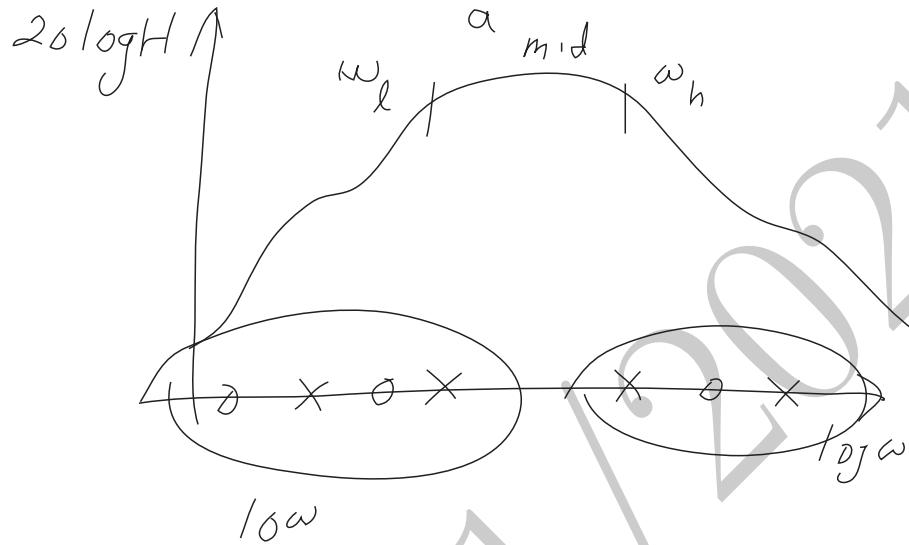


Figure 4.2: A Band Pass Transfer Function.

low-pass amplifiers, which is an amplifier whose gain in the band of interest (before it starts to drop off) is equal to its dc gain (Figure 4.1). An example of such is a dc-coupled amplifier, such as an operational-amplifier (op-amp).

Considering (4.1) and (4.2), in the special case where the poles are real and not too close to each other the pole frequencies can be approximated as

$$p_1 = -1/b_1 \quad (4.3a)$$

$$p_2 = -b_1/b_2 \quad (4.3b)$$

⋮

which is often referred to as the *dominant pole approximation*. We will see later when this approximation is accurate and how it is used.

The form used in (4.2) is not unique. In the case of a band-pass amplifier, the poles and the zeros can usually be divided into two groups: those occurring close to and below the low -3dB cut-off frequency,  $\omega_l$  and those that fall near and above the high -3dB cut-off frequency,  $\omega_h$ , as depicted in Figure 4.2. In this case, the gain we are most interested in is not the low frequency gain,  $a_0$ , which can be very low or even zero. Rather we care about the mid-band gain,  $a_{mid}$ , shown in Figure 4.2<sup>6</sup>. If there are  $k$  poles and  $k$  zeros below the mid-band<sup>7</sup>,

<sup>6</sup>Not every transfer function has a well-defined mid-band gain. For example, if there is not breakpoint between the lower and upper set of poles and zeros, there will not be a broad range frequencies with a constant gain. Nevertheless, a properly designed broadband amplifier often does demonstrate a region with a relatively constant gain.

<sup>7</sup>The number of poles and zeros below the mid-band *must* be equal to have a flat mid-band

(4.2) can be reordered as:

$$H(s) = \frac{(1 - \frac{z_1}{s}) \dots (1 - \frac{z_k}{s})}{(1 - \frac{p_1}{s}) \dots (1 - \frac{p_k}{s})} \cdot a_{mid} \cdot \frac{(1 - \frac{s}{z_{k+1}}) \dots (1 - \frac{s}{z_m})}{(1 - \frac{s}{p_{k+1}}) \dots (1 - \frac{s}{p_n})} \quad (4.4)$$

where the terms to the left of the mid-band gain<sup>8</sup>,  $a_{mid}$ , are written in the *inverse poles* and *inverse zeros* format. This representation is helpful when we try to separate the effect of the poles and zeros affecting  $\omega_l$  from those controlling  $\omega_h$ .

As mentioned in our discussion of the nodal analysis in Chapter 2, the denominator of (4.1) is proportional to the determinant of the  $Y$  matrix of circuit with a scalar frequency-independent proportionality constant. The order of the denominator,  $n$ , determines the number of natural frequencies of the system and is equal to the number of *independent* energy storage elements, as we will show later in Section 4.6. The number of independent energy-storage elements is also equal to the maximum number of independent initial conditions (capacitor voltages and inductor currents) that can be set.

Since the  $Y$  matrix is independent of our definition of the input or the output variables, its determinant and thereby its roots are also independent of the choice of the input and output variables and is an intrinsic characteristic of the circuit. This means that the pole frequencies (which represent the natural modes of the circuit) are a global property of the circuit independent of the point of observation<sup>9</sup>.

On the contrary, the zeros of the transfer function (i.e., the roots of the numerator of (4.1)) *do* depend on the choice of the input and output, as can be seen from the nodal analysis discussions of Chapter 2. While it is possible to answer what the poles of a circuit are without knowing what the input and output variables are, it is meaningless to ask the same question about the zeros, as they assume different values for different choices of the input or the output.

Knowing the poles and zeros of an LTI system, we can predict its dynamics. In the following sections we progressively develop a method to determine these initially for a first-order system and then for a broader class of systems.

## 4.2 System with a Single Energy Storage Element

Online YouTube lectures:

### Time- and transfer-constants in 1st order system

---

region. This is because every pole introduces a drop at the rate of  $-20dB/dec$  and each zero introduces an increase of  $+20dB/dec$  in the amplitude response.

<sup>8</sup>It is easy to verify that  $a_{mid}$  and  $a_0$  are related through:

$$a_0 = a_{mid} \cdot \frac{z_1 z_2 \dots z_k}{p_1 p_2 \dots p_k} \quad (4.5)$$

<sup>9</sup>The implicit assumption here is that these modes are *observable* in the linear system theory sense.

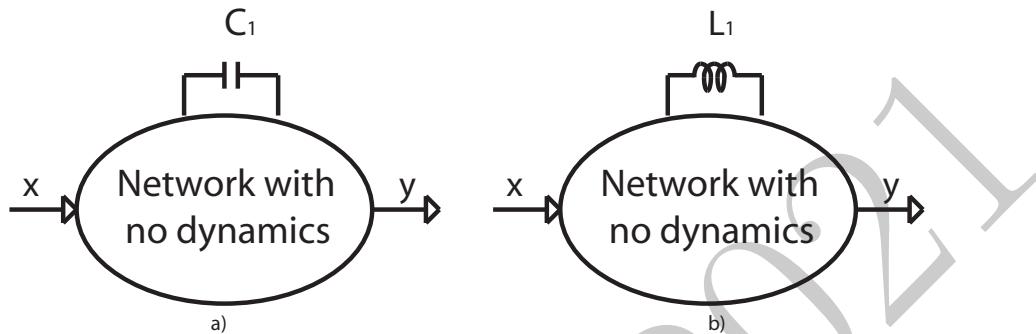


Figure 4.3: A first order system with a) a capacitor as the energy storing element, b) an inductor.

Let us consider an LTI circuit which has only one energy-storage element with an input  $u_i$  and an output  $u_o$ , as shown in Figure 4.3a and b for a system with a capacitor or an inductor, respectively. In general,  $u_i$  and  $u_o$  could be the voltages and/or currents of any two ports, including the current and voltage of the same port. For example, when  $u_i$  is an input voltage driven by a voltage source and  $u_o$  is the voltage of another node in the circuit, the transfer function,  $H(s) \equiv v_o(s)/v_i(s)$ , would correspond to a voltage gain. On the other hand, if the input,  $u_i$ , is the current of a current source driving a given port of the circuit, while the output,  $u_o$ , is the voltage across the *same* port, the transfer function,  $Z(s) \equiv v_1(s)/i_1(s)$  would correspond to the driving-point impedance looking into that port<sup>10</sup>.

Although the circuits of Figure 4.3 includes only one reactive element,  $C_1$  or  $L_1$ , the network in the box can be quite complex with any number of frequency-independent elements, such as resistors and dependent current and voltage source. In fact all the circuit we have analyzed in Chapter 3 would fall in this category.

A circuit with one reactive element will have exactly one pole and one zero, which can occur at any two frequencies between and including  $-\infty$  and  $+\infty$ . (It is customary to say there is “no zero” when it is at infinity.) For this first

<sup>10</sup>One has to be careful here with the choice of the stimulus and the output. If a node is excited with a current source and the voltage across that node is measured, then the quantity measured is the impedance,  $Z(s)$ . On the other hand, if the same port is excited by a voltage source and the current is the output variable, the calculated transfer function is the admittance  $Y(s)$ . Although in the end we must have  $Z(s) = 1/Y(s)$ , one should be very careful to keep things consistent, as the poles of  $Z(s)$  are the zeros of  $Y(s)$  and vice versa. As we will see the procedure used to calculate the poles and zeros is different, and relies on nulling the independent source, which means a short-circuit for a voltage source and an open-circuit for the current source. If we fail to keep things consistent we will obtain erroneous results in the end. We will elaborate on this via several examples later.



Figure 4.4: The equivalent circuits to Figure 4.3 with no input present for a)capacitor, b)inductor.

order system, the general transfer function of (4.1) reduces to:

$$H(s) = \frac{a_0 + a_1 s}{1 + b_1 s} \quad (4.6)$$

where  $a_0$  is the low-frequency transfer function. The pole will be at  $p = -1/b_1$  corresponding to a pole time constant of  $\tau \equiv b_1$ . The zero occurs at  $z = -a_0/a_1$ . For a zero occurring at infinity,  $a_1 = 0$ , and for a zero at the origin,  $a_0 = 0$ .

We will designate the value of the transfer function when the reactive element (or in general all reactive elements) is (are) zero valued ( $C = 0$ , i.e., open circuited capacitor and/or  $L = 0$ , i.e., short circuited inductor) as  $H^0$ . This is the same as the the low frequency transfer function since setting every reactive elements to zero removes any frequency dependence from the circuit, i.e.,

$$a_0 = H^0 \quad (4.7)$$

It is a well-known result that the time constant,  $\tau$ , of a first-order circuit with a capacitor,  $C_1$ , is  $R^0 C_1$ , where  $R^0$  is the resistance seen across the two nodes where the capacitor is connected to without the capacitor (when  $C_1 = 0$ ) and *all* the independent sources, *including the input source* are *nulled*. We usually say that this is the impedance “seen” by the capacitor. As discussed earlier, nulling these sources means replacing an independent voltage source with a short circuit and an independent current source with an open-circuit. In this case, the circuit of Figure 4.3a simply reduces to the parallel combination of capacitor,  $C_1$ , and the low frequency resistance seen by it,  $R^0$ , as shown in Figure 4.4a. This result in a pole time constant of

$$\tau \equiv R^0 C_1 \quad (4.8)$$

where the superscript zero in  $R^0$  indicates that the independent sources and the energy-storing element are at their zero values. We will consider the case of an inductive energy-storing element later.

#### ▼ Derivation ▼

In the transfer function of (4.6), the capacitor is the only frequency dependent element and thus  $C$  is the sole term accompanied by the complex frequency,  $s$ . The impedance of the capacitor  $C_1$  is simply  $1/C_1 s$ . We simply notice that the capacitance,  $C_1$ , and the complex frequency,  $s$ , always appear together as a product, so the transfer function of (4.6) can be unambiguously written as:

$$H(s) = \frac{a_0 + \alpha_1^1 C_1 s}{1 + \beta_1^1 C_1 s} \quad (4.9)$$

where  $\beta_1^1$  has units of  $[\Omega]$  and  $\alpha_1^1$  units of  $\Omega$  times the units of the transfer function itself (e.g., if the transfer function in question is an impedance, its units will be  $[\Omega^2]$  in this case)<sup>11</sup>. Note that the superscript “1” is used as an index and not an exponent. Combining (4.7)-(4.9) we have

$$\beta_1^1 = R^0 \quad (4.10)$$

and hence

$$b_1 = R^0 C_1 = \tau \quad (4.11)$$

Now, we can focus our attention on the zero. This time assume that the capacitor is at its infinite value, i.e.,  $C_1 \rightarrow \infty$ . For a capacitor this is equivalent to having it replaced with a short circuit. For  $C_1 \rightarrow \infty$ , the second terms in the numerator and the denominator of the transfer function of (4.9) dominate and hence it reduces to:

$$H^1 \equiv H|_{C_1 \rightarrow \infty} = \frac{\alpha_1^1}{\beta_1^1} \quad (4.12)$$

where  $H^1$  is the frequency-independent *transfer constant* from the input,  $u_i$ , to the output  $u_o$  with the capacitor,  $C_1$  at its *infinite value*, namely, short circuited. This is simply another low-frequency gain calculation that can be readily performed using the techniques of Chapter 3. Note that in general this is a different transfer constant from  $H^0$  which is the low frequency transfer constant with the capacitor being zero (open circuited).

A comment on notation is in place here. From this point on, unless explicitly stated otherwise, the superscript index(es) refer to the index(es) of the energy-storage elements that are *infinite valued* (i.e., shorted capacitors and/or opened inductors). A superscript of 0 corresponds to the case where *no* energy-storing element is infinite valued, or equivalently *all* energy-storing elements are zero-valued.

Considering (4.10)-(4.12) and comparing (4.6) to (4.9), we easily determine  $a_1$  to be:

$$a_1 = \alpha_1^1 C_1 = R^0 C_1 H^1 = \tau H^1 \quad (4.13)$$

where  $\tau$  is simply the pole time constant defined in (4.8).

Equations (4.7), (4.11), and (4.13) provide a straightforward procedure to determine the transfer function in terms of the resistance seen by the capacitor

---

<sup>11</sup>The  $\alpha$  and  $\beta$  in this derivation has nothing to do with BJT's  $\alpha$  and  $\beta$  and are rather intermediate variables used in this and some of the subsequent derivations.

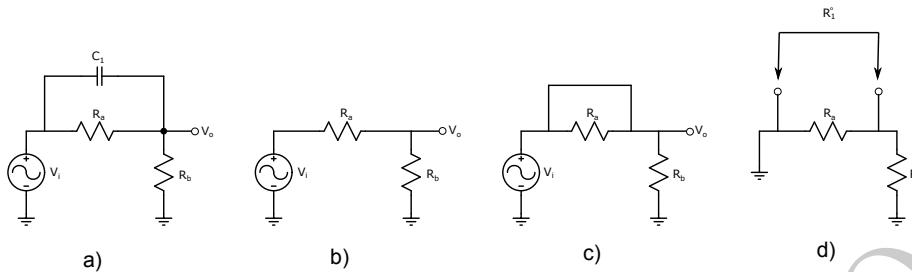


Figure 4.5: a) An  $RC$  high-pass filter. b) Low frequency model for  $H^0$  calculation, c) the low frequency model for  $H^1$  calculation, d) the low frequency circuit for  $\tau_0$  calculation.

with the input nulled ( $R^0$ ) and the values of the transfer function at dc with the capacitor,  $C_1$ , open- and short-circuited, namely, the two *transfer constants*:  $H^0$  and  $H^1$ , respectively. We can write this transfer function as:

$$H(s) = \frac{H^0 + \tau H^1 s}{1 + \tau s} \quad (4.14)$$

where  $\tau$  is calculated using (4.8) for a capacitor.

If the energy-storage element is an inductor,  $L_1$ , (Figures 4.3b and 4.4b) a similar argument leads to (4.14). In this case,  $H^0$  is the transfer constant calculated for  $L_1$  zero-valued (short-circuited) and  $H^1$  is another transfer constant calculated for  $L_1$  infinite-valued (open-circuited). The inductor sees the same resistance as in the case of a capacitor, namely  $R^0$ , and hence the time constant  $\tau$  is given by

$$\tau = \frac{L_1}{R^0} \quad (4.15)$$

As can be seen for a single energy-storing element, (4.14) provides the exact transfer function of the system, in terms of three low-frequency calculations.

Now let us look at a few applications of (4.14). The first two examples are somewhat trivial to familiarize the reader with the procedure. In some of the examples, we replace the  $H$  and its associated terms with the appropriate letter (e.g.,  $Z$  for impedance and  $Y$  for admittance.)

**Example 4.2.1 (High-Pass RC)** Consider the simple  $RC$  circuit driven by an ideal voltage source, depicted in Figure 4.5a. The low frequency gain is simply given by setting  $C_1 = 0$ , which results in a voltage divider illustrated in Figure 4.5b, hence:

$$H^0 = \frac{R_2}{R_1 + R_2}$$

To calculate the transfer constant  $H^1$ , we set  $C_1$  to its infinite value, i.e., short circuit it. It is clear from Figure 4.5c that in this case, we have

$$H^1 = 1$$

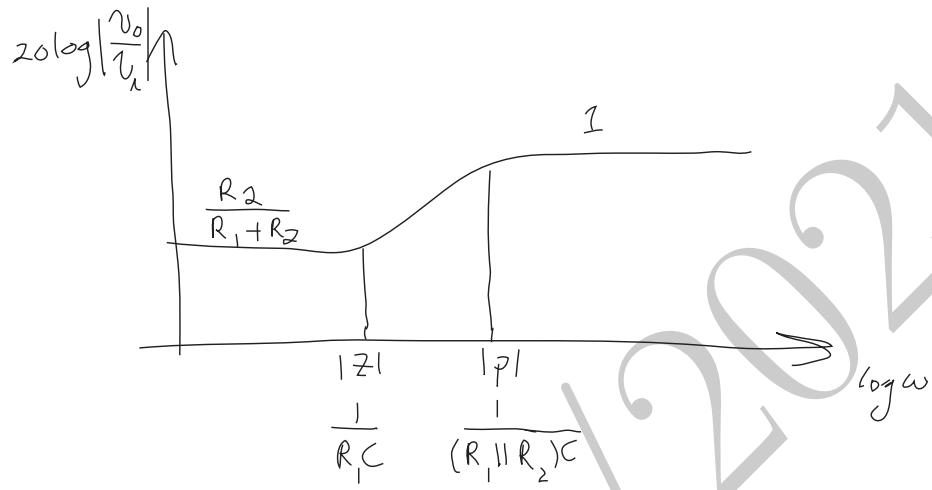


Figure 4.6: The transfer function of the  $RC$  high-pass circuit of Figure 4.5

since the input and output nodes are short circuited together. Finally, to calculate  $\tau$  we need to determine,  $R^0$ , the resistance seen by  $C_1$  when input voltage source is replaced with a short-circuit to ground (independent source nulled). In this case,  $R_1$  and  $R_2$  will be in parallel, as seen in Figure 4.5d, resulting in

$$\tau = R^0 C = (R_1 \parallel R_2) C$$

Using (4.14), we obtain

$$H(s) = \frac{R_2}{R_1 + R_2} \cdot \frac{1 + R_1 C s}{1 + (R_1 \parallel R_2) C s}$$

with  $p = -1/(R_1 \parallel R_2)C$  and  $z = -1/R_1 C$ . The transfer function of this circuit is shown in Figure 4.6.

Here is another simple example:

**Example 4.2.2 (High Pass LR)** The low-frequency voltage transfer function of the  $RL$  circuit of Figure 4.7a can be easily calculated noting,

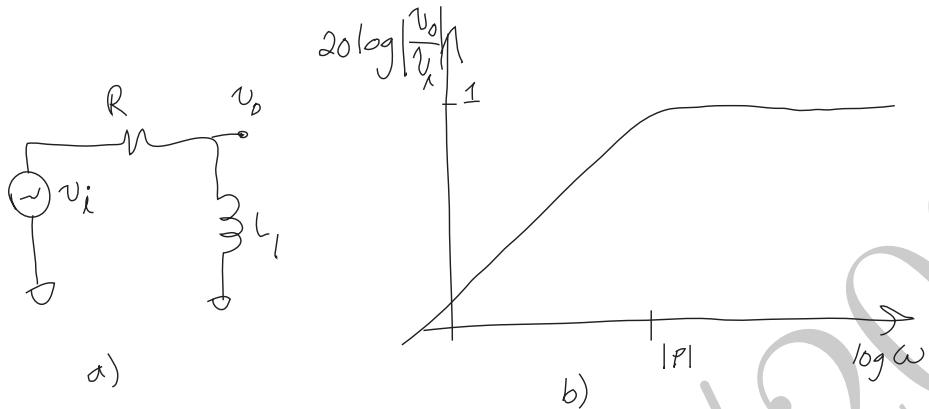
$$H^0 = 0$$

because the inductor shorts the output node to the ground at low frequency. For  $L_1 \rightarrow \infty$ , the inductor is an open-circuit and hence  $v_o = v_i$ , hence

$$H^1 = 1$$

To calculate  $\tau$ , we null the input voltage source (short). In this case, the inductor sees the resistor  $R_1$  across itself, so

$$\tau = \frac{L_1}{R^0} = \frac{L_1}{R_1}$$

Figure 4.7: a) An  $RL$  low-pass filter b)Its high-pass transfer function.

Using (4.14) we have

$$H(s) = \frac{\frac{L_1}{R_1}s}{1 + \frac{L_1}{R_1}s} = \frac{1}{1 + \frac{R_1}{L_1}\frac{1}{s}}$$

with  $p = -R_1/L_1$  and  $z = 0$  (zero at the origin). This combination, as we discussed before is better presented as an inverse pole, as is shown on the right-hand side of the above equation. The transfer function of this circuit is shown in Figure 4.7b.

At this point it may appear that the above approach has made the calculation more complicated compared to the standard application of KCL and KVL or nodal analysis. However, as we will see in the rest of this Chapter, the above approach is more modular and scalable when the size of the problem and the number of energy-storing elements increases. Unlike nodal analysis, it can provide partial results and useful information without having to carry the analysis to the end.

Online YouTube lectures:  
[Time- and transfer-constants in 1st order system](#)

**Example 4.2.3 (Source Follower: Voltage Gain)** Consider the source follower of Figure 4.8a, with  $C_\pi$  as the only capacitance in the circuit. The equivalent small-signal model is illustrated in Figure 4.8b.

The zero value transfer constant is

$$H^0 = \frac{v_o}{v_i} = \frac{R_2}{r_m + R_2} = \frac{g_m R_2}{1 + g_m R_2} = a_0 \quad (4.16)$$

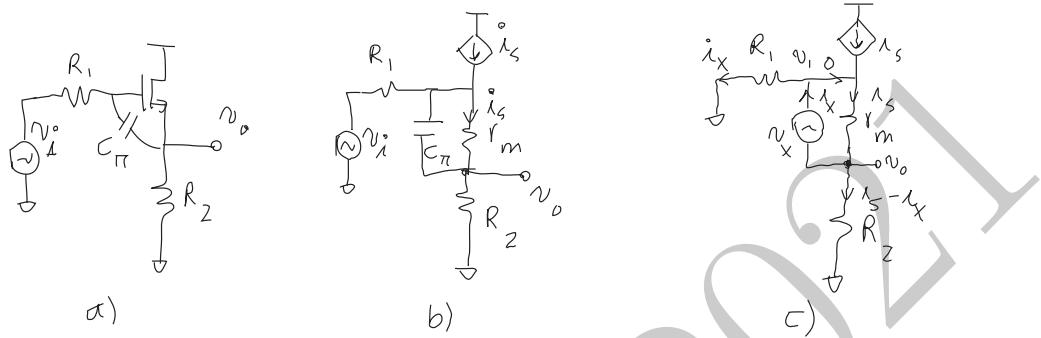


Figure 4.8: a) A source follower stage with a source resistance  $R_1$ , a load resistance,  $R_2$ , and a gate-source capacitance,  $C_\pi$ , b) Its small-signal model, c) The test source applied to calculate  $R^0$ .

The infinite-value ( $C_\pi \rightarrow \infty$ ) transfer constant can be easily calculated by short circuiting  $C_\pi$  which results in a voltage divider between  $R_1$  and  $R_2$ ,

$$H^\pi = \frac{R_2}{R_1 + R_2}$$

The resistance seen by  $C_\pi$  can be calculated by using the T-model shown in Figure 4.8c, applying a test voltage source,  $v_x$  and measuring the current  $i_x$ . We can easily see that  $v_o$  is the total current entering resistance  $R_2$  times  $R_2$  itself, i.e.,

$$v_o = (i_s - i_x)R_2 = (g_m v_x - i_x)R_2$$

Noting that,  $v_1 = v_o + v_x$  and the fact that the current  $i_x$  directly flows into  $R_1$ , we can write

$$R_1 i_x = v_1 = v_o + v_x = g_m R_2 v_x - i_x R_2 + v_x = (1 + g_m R_2) v_x - R_2 i_x$$

Reordering, we find that

$$R^0 = \frac{v_x}{i_x} = \frac{R_1 + R_2}{1 + g_m R_2} \quad (4.17)$$

that leads to the time-constant:

$$\tau = R^0 C_\pi = r_m C_\pi \cdot \frac{R_1 + R_2}{r_m + R_2} \quad (4.18)$$

Using these results in (4.14), we simply obtain:

$$H(s) = \frac{v_o(s)}{v_i(s)} = H^0 \cdot \frac{1 + \frac{H^\pi}{H^0} \tau s}{1 + \tau s} = \frac{R_2}{R_2 + r_m} \cdot \frac{1 + \frac{C_\pi}{g_m} s}{1 + R^0 C_\pi s}$$

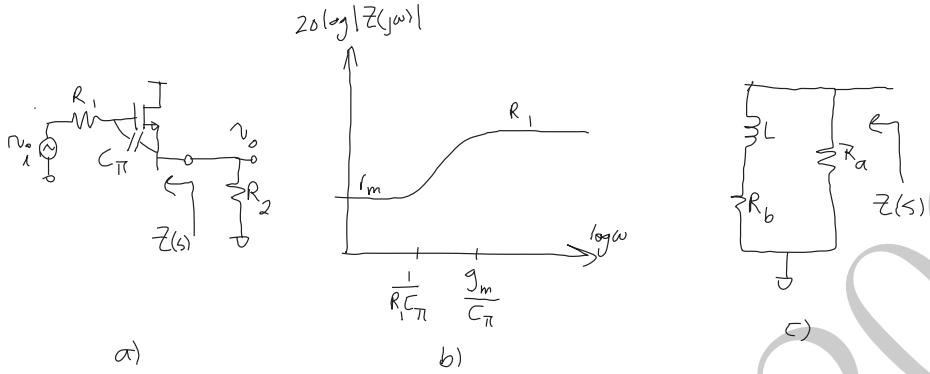


Figure 4.9: a) Output impedance of source follower with a source resistance,  $R_1$ , and a gate-source capacitance,  $C_\pi$ , b)  $|Z(j\omega)|$  vs. frequency for  $R_1 > r_m$ , c) the output equivalent circuit.

with a left-half plane pole at  $p = -1/R^0 C_\pi$  and a left-half plane zero at  $z = -g_m/C_\pi$ . Note that zero's frequency calculated is equal to the cut-off frequency,  $\omega_T$  of the transistor calculated in (1.61) assuming that  $C_\mu$  is zero. If  $R_1$  and  $R_2$  are comparable (or if  $R_2 \gg R_1$ ), the pole frequency will also be on the same order of magnitude as  $-g_m/C_\pi$  and  $\omega_T$ . Hence we have a pole-zero pair at relatively high frequencies. Once we take  $C_\mu$  into account in Example 4.5.2, we will see that for the values of the input resistor  $R_1$  that are not too small, there is usually another pole that occurs at a lower frequency than the above pole and zero and hence dominates the response. One way to explain why the resistance  $R^0$  seen by  $C_\pi$  is reduced from  $R_1 + R_2$  by a factor of  $1 + g_m R_2$  will be given in Section 4.2.3 on page 143.

**Example 4.2.4 (Source Follower: Output Impedance)** Now we determine the output impedance of the source follower of Figure 4.9a. As we can see, the output impedance is the parallel combination of the intrinsic output impedance  $Z(s)$  and the resistance  $R_2$ . We will focus our attention on  $Z(s)$  because once we calculate it, the extrinsic output impedance is simply  $Z(s) \parallel R_2$ .

Assuming  $R_2 \rightarrow \infty$ , the zero-value transfer constant is the intrinsic output resistance  $Z^0 = r_m$ . The infinite-value ( $C_\pi \rightarrow \infty$ ) output impedance is simply  $Z^\infty = R_1$  when the capacitor is short circuited. Finally for  $R_2 \rightarrow \infty$ , the resistance seen by  $C_\pi$  is simply  $R^0 = r_m$  and therefore  $\tau = r_m C_\pi$ . Combining these three results and using (4.14), we simply have:

$$Z(s) = r_m \cdot \frac{1 + R_1 C_\pi s}{1 + r_m C_\pi s} \quad (4.19)$$

we note that the pole occurs at  $p = -g_m/C_\pi$  which is slightly higher than the transistor cut-off frequency,  $\omega_T$ . The zero occurs at  $z = -1/R_1 C_\pi$  which occurs at a lower frequency than the pole as long as  $R_1 > r_m$  which is usually the

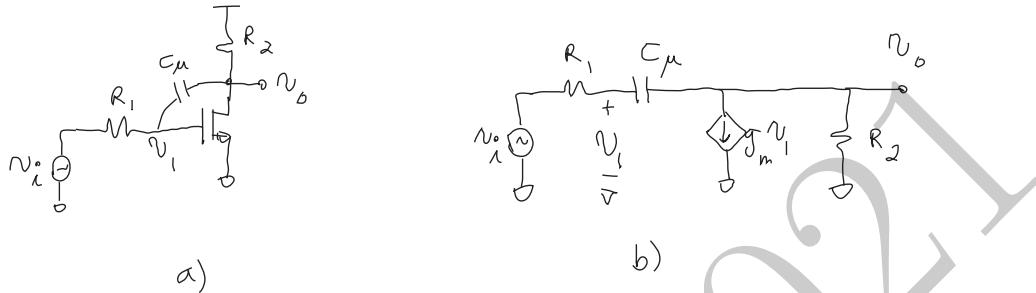


Figure 4.10: a) A common-source amplifying stage with only  $C_\mu$  considered, b) its small-signal equivalent model.

case. If this is the case, the impedance will increase with the frequency between the zero and the pole frequencies, as shown in Figure 4.9b, and thus behaves as an inductor. This behavior is similar to that of the equivalent circuit shown in Figure 4.9c. We need to choose  $L$ ,  $R_a$ , and  $R_b$  in such a way that the impedance of the equivalent circuit of Figure 4.9c matches (4.19). We notice that at infinite frequency ( $L$  open), we must have  $R_a = R_1$ , and for  $s = 0$  (low frequency), we must have  $R_a \parallel R_b = r_m$ . Solving for  $R_a$  we have:

$$\begin{aligned} R_a &= R_1 \\ R_b &= r_m \cdot \frac{R_1}{R_1 - r_m} \approx r_m \end{aligned}$$

Approximating  $R_b$  as  $r_m$  is valid as long as  $R_1 \gg r_m$ . Note that  $R_b$  becomes negative for  $R_1 < r_m$  which is consistent with the earlier observation that the output is inductive only for  $R_1 > r_m$  beyond which the equivalent model of Figure 4.9c is not valid anymore.

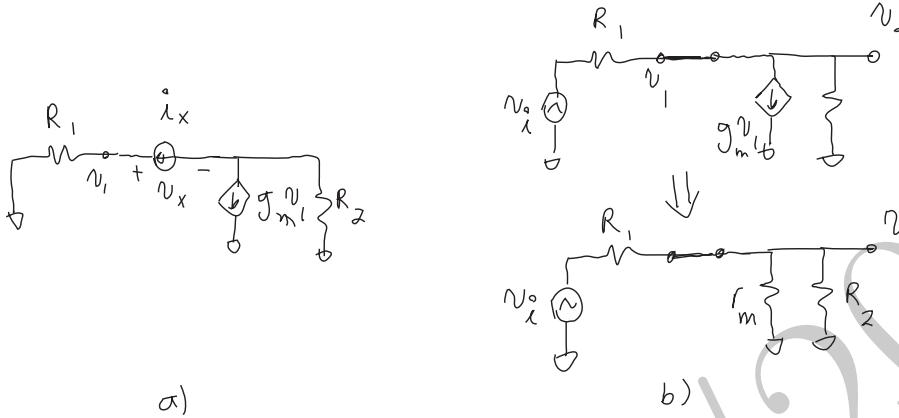
At this point, the impedance of the inductive equivalent circuit of Figure 4.9c is given by

$$Z_{eq}(s) \approx r_m \frac{1 + \frac{L}{r_m} s}{1 + \frac{L}{R_1} s}$$

For this impedance to match the output impedance of the source-follower stage we should have

$$L = r_m R_1 C_\pi = \frac{R_1 C_\pi}{g_m}$$

**Example 4.2.5 (Common-Source: Voltage Gain)** Considering a common-source inverting amplifier with a transconductance,  $g_m$ , driving a load resistor of  $R_2$  being driven by a voltage source  $v_{in}$  with a source resistance  $R_1$ , as shown in Figure 4.10a and its small-signal equivalent model in Figure 4.10b. Let us assume that the only reactive element in this circuit is the capacitor  $C_\mu$  connected between the input and output of the amplifier. The zero-value transfer constant

Figure 4.11: Calculation of a)  $R^0$ , b)  $H^\mu$ .

is easily calculated to be:

$$H^0 = \frac{v_o}{v_i} = -g_m R_2$$

To determine the resistance seen by  $C_\mu$  we apply a test current,  $i_x$ , in its place and determine the voltage drop, as shown in Figure 4.11a. This can easily be seen to be given by

$$v_x = v_1 - v_o = v_1 - (-i_x - g_m v_1) R_2 = v_1 (1 + g_m R_2) + i_x R_2$$

Noting that  $v_1 = i_x R_1$ , we have<sup>12</sup>

$$R^0 = \frac{v_x}{i_x} = R_1 + R_2 + g_m R_1 R_2 \quad (4.20)$$

Therefore, the time constant is

$$\tau = R^0 C_\mu = (R_1 + R_2 + g_m R_1 R_2) C_\mu \quad (4.21)$$

For  $C_\mu \rightarrow \infty$ , the capacitor is short-circuited and hence  $v_o = v_1$ . As a result, the circuit reduces to the one shown in Figure 4.11b. Note that now the dependent current source,  $g_m v_1$ , is proportional to its own voltage,  $v_1$ , and thus can be replaced with a resistance  $r_m = 1/g_m$ . The voltage gain is simply given by the resistive divider ratio, i.e.,

$$H^\mu = \frac{r_m \parallel R_2}{R_1 + r_m \parallel R_2} = \frac{R_2}{R_1 + R_2 + g_m R_1 R_2} = \frac{R_2}{R^0} \quad (4.22)$$

Hence the voltage transfer function can be determined from (4.14).

$$H(s) = \frac{v_o(s)}{v_i(s)} = H^0 \cdot \frac{1 + \frac{H^\mu}{H^0} \tau s}{1 + \tau s} = -g_m R_2 \cdot \frac{1 - \frac{C_\mu}{g_m} s}{1 + R^0 C_\mu s}$$

<sup>12</sup>When  $|a_0| = g_m R_L \gg 1$ , we have  $R^0 \approx g_m R_1 R_2$

with a LHP pole at  $p = -1/R^0 C_\mu$  and a RHP zero at  $z = g_m/C_\mu$ . Note that zero's frequency calculated is equal to the cut-off frequency,  $\omega_T$  of the transistor calculated in (1.61) assuming that  $C_\pi$  is zero. We will see later that when both  $C_\pi$  and  $C_\mu$  are included in the calculations, the zero frequency remains the same and this is even higher than  $\omega_T = g_m/(C_\pi + C_\mu)$ .

It is noteworthy that in this example,  $H^0$  and  $H^1$  have *opposite* signs which results in a right-half plane (RHP) zero in the transfer function. In general, the relative magnitude and size of  $H^0$  and  $H^1$  can provide us with useful information about the properties of the zero.

We will see more formally later that in general a simple test to determine if there are *any* zeros in the transfer function of a system with multiple energy-storing elements is whether shorting of *any* capacitor or opening of any inductor in the circuit results in a non-zero low-frequency transfer function.

#### 4.2.1 Zeros in a First-Order System

Online YouTube lectures:

[Zero/pole in 1st order system, step response, undershoot and overshoot](#)

For a system with a single energy-storing element,  $C_1$  (or  $L_1$ ), looking at (4.14), we can easily obtain the following relation between the pole and the zero in terms of the ratio of two transfer constants:

$$z = \frac{H^0}{H^1} \cdot p \quad (4.23)$$

This expression is sufficient to evaluate the relative position of the zero with respect to the pole. For instance, it is clear from (4.23) that if the infinite- and zero-value transfer constants have opposite signs, the poles and the zero will be on two opposite half-planes. In stable systems where the pole is in the LHP, the zero will be on the RHP for opposite polarities of  $H^0$  and  $H^1$ , as in Example 4.2.5. On the other hand if  $H^0$  and  $H^1$  have the same polarity, the pole and the zero will be both on the LHP resulting in the so-called minimum phase system, similar to Example 4.2.3.

While the sign of  $H^0/H^1$  determines whether or not the pole and the zero are on the same side of the  $j\omega$  axis, its magnitude determines which one occurs at a lower frequency. As evident from (4.23), the zero happens first (at a lower frequency than the pole) when  $|H^0/H^1| < 1$ , namely,  $|z| < |p|$ . Alternatively, the pole will happen before the zero ( $|p| < |z|$ ), for  $|H^0/H^1| > 1$ .

The magnitude and the sign of  $H^0/H^1$  provide useful information about the relative location of the pole and the zero in a first order system. This can almost always be done by inspection because we only need to know the relative size and magnitude of  $H^0$  and  $H^1$ . This is summarized in Table 4.1 and shown using magnitude and phase plots of the four different situations in Figure 4.12.

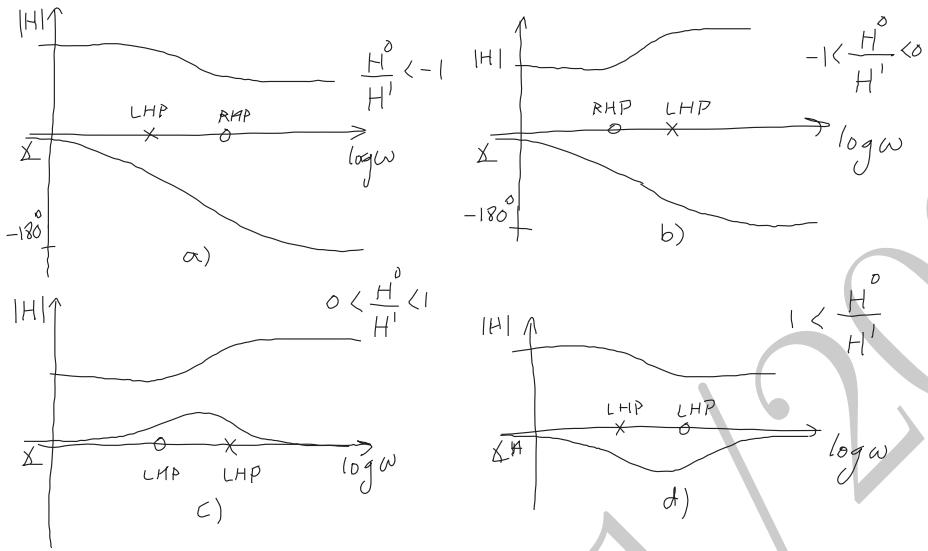


Figure 4.12: The magnitude and phase plots for a)  $|H^0/H^1| < -1$ , b)  $-1 < |H^0/H^1| < 1$ , c)  $|H^0/H^1| > 1$ , and d)  $|H^0/H^1| > 1$

#### 4.2.2 The Time-Domain Response of a First Order System

The step response of a first order system with the transfer function given by (4.14) can be easily calculate noting that (4.14) can be written as the sum of two first order transfer functions: one with a single pole (first order low-pass) and the other with a single *inverse* pole (first order high-pass), i.e.,

$$H(s) = \frac{H^0}{1 + \tau s} + \frac{H^1}{1 + \frac{1}{\tau s}} \quad (4.24)$$

Hence, the step response is the sum of the step responses of the first order low- and the high-pass systems weighted by  $H^0$  and  $H^1$ , respectively, which can be

Table 4.1: Relative position of the pole and zero in a first-order system as a function of  $H^0$  and  $H^1$ .

	$ \frac{H^0}{H^1}  < 1$	$ \frac{H^0}{H^1}  > 1$
$\frac{H^0}{H^1} > 0$	$ z  <  p $ Same Half Plane	$ z  >  p $ Same Half Plane
$\frac{H^0}{H^1} < 0$	$ z  <  p $ Opposite Half Plane	$ z  >  p $ Opposite Half Plane

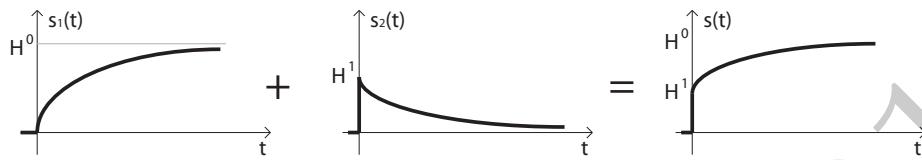


Figure 4.13: The step response of a first order system decomposed as the sum of the step response of a first-order low- and high-pass systems.

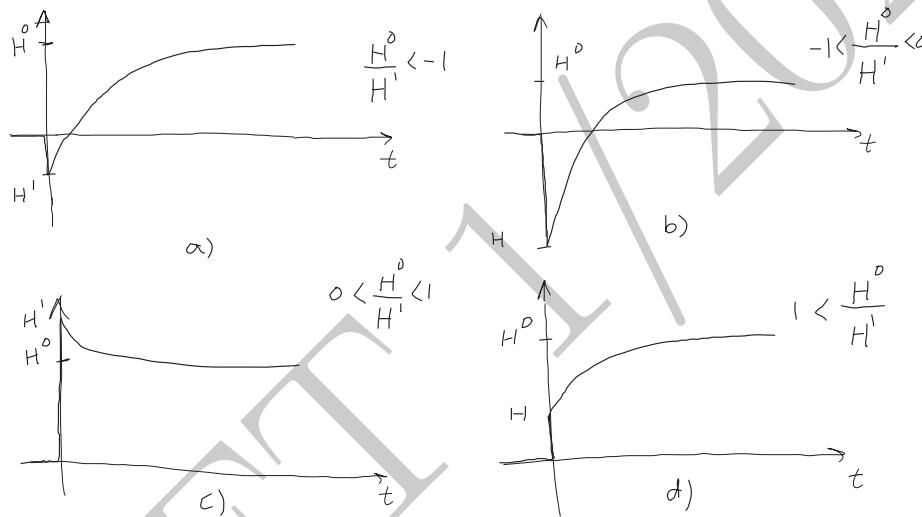


Figure 4.14: The step response of a first order system for a)  $H^0/H^1 < -1$ , b)  $-1 < H^0/H^1 < 0$ , c)  $0 < H^0/H^1 < 1$ , and d)  $1 < H^0/H^1$

stated as,

$$s(t) = H^0(1 - e^{-t/\tau})u(t) + H^1e^{-t/\tau}u(t) \quad (4.25)$$

where  $u(t)$  is the unit step. This decomposition is shown in Figure 4.13. Since both responses have the same time constant,  $\tau$ , the overall response would be an exponential with an *initial* value,  $H^1$ , and a *final* value,  $H^0$ , with a time constant,  $\tau$ , as shown in Figure 4.13. Again, as in the case of the frequency response discussed in the previous subsection, the relative size and polarities of  $H^0$  and  $H^1$  determines the general behavior of the response.

When  $H^0$  and  $H^1$  have opposite polarities, the low- and high pass responses will go in different directions resulting in an undershoot<sup>13</sup>, as shown in Figure 4.14a and b. On the other hand, when  $H^0$  and  $H^1$  have the same polarities, but  $0 < H^0/H^1 < 1$ , the step response's initial value ( $H^1$ ) is greater than its final

<sup>13</sup>Another way to see this is by noting that the initial value and the final values in Figure 4.13 will have opposite signs.

value ( $H^0$ ) and hence, there will be an overshoot, as shown in Figure 4.14c. For  $1 < H^0/H^1$ , the output instantaneously jumps to  $H^1$  at  $t = 0$  and then follows an exponential to its final value, as shown in Figure 4.14d. Note that in the special case, where  $H^0 = H^1$ , the pole and the inverse pole responses add to form a perfect step, i.e., the pole and the zero cancel each other.

We will have more discussions about the zeros and their effect on the response of the system in the Section 4.4.2.

### 4.2.3 Miller Effect

Online YouTube lectures:

[Miller multiplication and effect, input impedance of common-source](#)

Let us consider the input impedance of the common-source stage as an opening to our discussion on the Miller Effect:

**Example 4.2.6 (Input Impedance: Miller Multiplication)** Considering the amplifier of the previous example, shown in Figure 4.10a, this time we will use (4.14) to determine its input impedance with  $C_\mu$  as the only energy-storing element. To do so, we note that the input impedance is the series combination of  $R_1$  and the impedance seen beyond it looking into the transistor,  $Z(s)$ , knowing that the input impedance will be  $R_1 + Z(s)$ . However, since the low frequency impedance is infinity in this case, it is more convenient to apply a voltage source and hence calculate the admittance,  $Y(s)$ . Note that once we choose a voltage source as the input variable and its current as the output variable, we are calculating the admittance and not the impedance. This is an important distinction to make. Although we can obtain  $Z(s)$  by inverting  $Y(s)$  once we obtain from the following calculation, using a voltage source and erroneously defining the transfer function as  $v_x/i_x$  (the ratio of the input to the output!) would result in an incorrect answer, as the poles of the  $Y(s)$  are the zeros of  $Z(s)$  and vice versa<sup>14</sup>.

In  $Y(s)$  calculation, the  $R^0$  is very easy to calculate with a voltage source nulled, node  $v_1$  is shorted to ground. Since  $v_1$  is zero, so is the dependent current source and hence  $R^0 = R_2$  and thus  $\tau = R_2 C_\mu$ . The transfer function is shown as  $Y(s)$  to emphasize that it is an admittance with units of siemens. Its low frequency value,  $a_0$ , denoted by  $Y^0$  is given by:

$$Y^0 = \frac{i_x}{v_x} \Big|_{C_\mu=0} = 0$$

<sup>14</sup>In this example, attempting to calculate  $Z(s)$  directly using a current source stimulus and defining the voltage as the output variable results in singularity since  $H^0 \rightarrow \infty$ . There are several ways to deal with this problem, the easiest of which is the above approach of calculating  $Y(s)$  instead and inverting it in the end. Another possibility is to introduce a resistor,  $R_3$  between  $v_1$  and ground, calculating  $Z(s)$  (now  $H^0$  does not blow up) and set  $R_3 \rightarrow \infty$  in the end. This is an example of the special case discussed in the footnote on Page 124.

since no current can flow through an open circuit. To determine  $Y^\mu$ , we short circuit  $C_\mu$  as in Figure 4.11b and obtain<sup>15</sup>

$$Y^\mu = \frac{i_x}{v_x}|_{C_\mu \rightarrow \infty} = g_m + \frac{1}{R_2} = \frac{1 + g_m R_2}{R_2}$$

Now we can calculate the admittance using (4.14)

$$Y(s) = \frac{i_x(s)}{v_x(s)} = \frac{Y^0 + Y^\mu \tau s}{1 + \tau s} = \frac{(1 + g_m R_2) C_\mu s}{1 + R_2 C_\mu s}$$

Now we can use this result to calculate the impedance,  $Z(s)$ , as follows:

$$\begin{aligned} Z(s) &\equiv \frac{1}{Y(s)} = \frac{1 + R_2 C_\mu s}{(1 + g_m R_2) C_\mu s} = \frac{1}{(1 + g_m R_2) C_\mu s} + \frac{R_2}{1 + g_m R_2} \\ &= \frac{1}{C_M s} + r_m \parallel R_2 \approx \frac{1}{C_M s} + r_m \end{aligned} \quad (4.26)$$

where  $C_M$  is defined as

$$C_M = (1 + g_m R_2) C_\mu \quad (4.27)$$

The approximation in the last step of (4.26) is justified if  $g_m R_2 \gg 1$ , i.e., the absolute value of the low-frequency gain is much greater than unity. We can generate an input equivalent circuit based on (4.26), as shown in Figure 4.15. Note that the input impedance consists of the extrinsic resistance  $R_1$ , the so-called Miller capacitance,  $C_M$ , and a series resistance  $r_m \parallel R_2$ . Unfortunately, this last resistance is often neglected in most treatments despite its importance in high frequency design (e.g., narrow-band RF design). For instance, if one tries to achieve an impedance match to the input using an input matching network, it is not possible to achieve a high frequency match to a real impedance without taking this resistor into account.

It is noteworthy that the capacitance seen at the input,  $C_M$  is  $1 + g_m R_2$  times greater than the actual capacitance,  $C_\mu$ . We note that the factor  $g_m R_2$  is negative voltage gain between the two terminals of the capacitor. So the equivalent input capacitance is simply

$$C_M = (1 + |A_v|) C_\mu = (1 - A_v) C_\mu \quad (4.28)$$

where  $A_v$  is simply the dc voltage gain across  $C_\mu$ . This effect is known as Miller multiplication and will be discussed in the following subsection.

---

<sup>15</sup>Note again that the notation  $A \parallel B$  designates one half of the harmonic mean of  $A$  and  $B$ , namely,  $A \parallel B \equiv AB/(A + B)$ . For resistors this would become the parallel combination and for conductances, it is the value of their series combination, i.e.,  $G_1 \parallel G_2 \equiv G_1 G_2 / (G_1 + G_2) = 1 / (R_1 + R_2)$ .

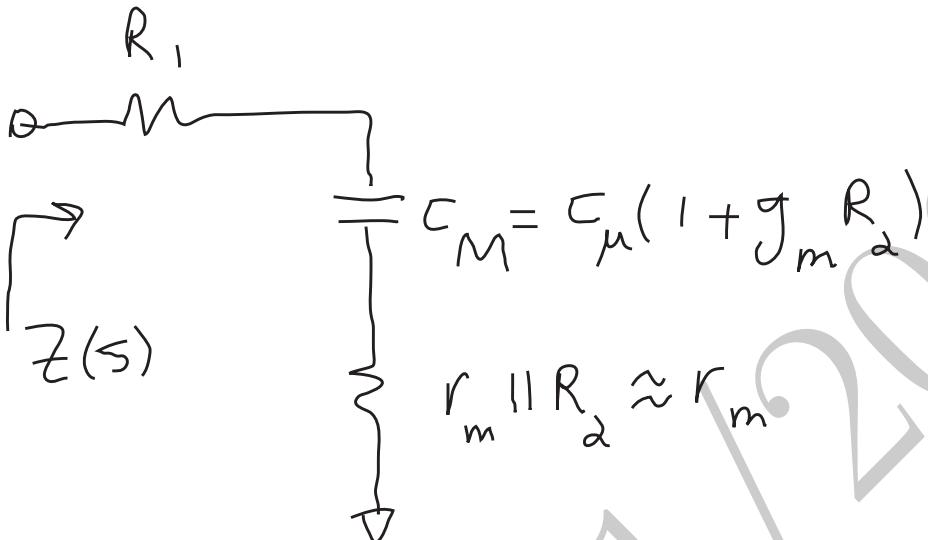


Figure 4.15: The equivalent circuit for the input of the common-source of Figure 4.10a.

### Miller Approximation

As we saw in the last example with an inverting voltage amplifier, the capacitance seen at the input is  $1 + |A_v|$  times greater than the capacitance appearing between the input and the output of the amplifier. This is known as the *Miller Effect* and was first explained in vacuum tubes by John Miller in a 1920 article.

Consider an ideal inverting voltage amplifier with an inverting gain  $A_v < 0$ . Being an ideal voltage amplifier its output impedance is zero<sup>16</sup>. Because of its inverting gain, an increase in the input voltage of value  $\Delta v$  results in a voltage decrease of exactly  $-|A_v|\Delta v = A_v\Delta v$  in the output, as shown in Figure 4.16. Consequently, a capacitor  $C$  connected between its input and output experiences a voltage excursion of  $(1 + |A_v|)\Delta v = (1 - A_v)\Delta v$ . This results in a current in the capacitor which is  $(1 + |A_v|)$  times greater than the current that would have flowed through it, had it been connected between the input and the ground, experiencing a voltage swing of just  $\Delta v$ . As far as the input current is concerned, this is equivalent to having a capacitor whose impedance is  $1 + |A_v|$  smaller than the original one. Since the capacitor's impedance is  $1/C_s$  and is thus inversely proportional to its capacitance, the equivalent capacitor at the input has to be  $1 + |A_v|$  larger than the the original capacitor, as seen in (4.28). Keep in mind that unlike this approach, the treatment given in Example 4.2.6 is exact even when the amplifier is not an ideal voltage amplifier.

Now consider the more general case of an impedance,  $Z$ , connected across

<sup>16</sup>The amplifier of Example 4.2.6 becomes an ideal voltage amplifier in the limit where  $g_m \rightarrow \infty$  and  $R_2 \rightarrow 0$  while the product remains constant at  $A_v = -g_m R_2$

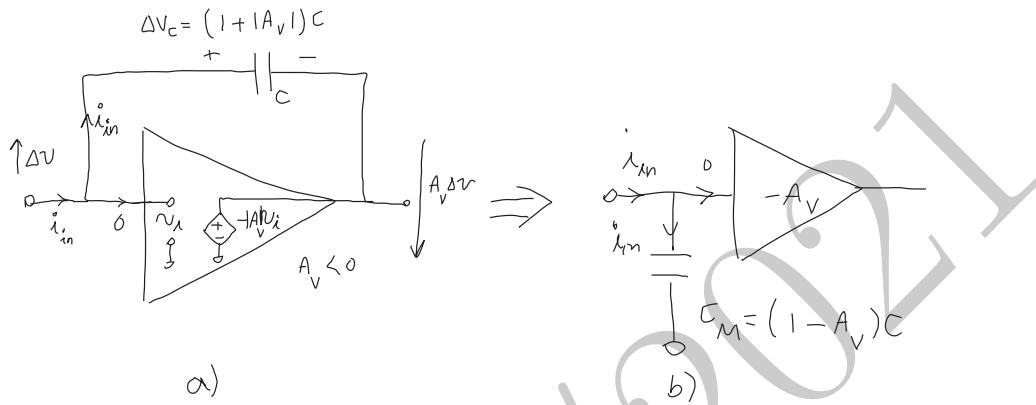


Figure 4.16: a) An ideal inverting amplifier with a capacitor between its input and output, b) its equivalent input capacitance.

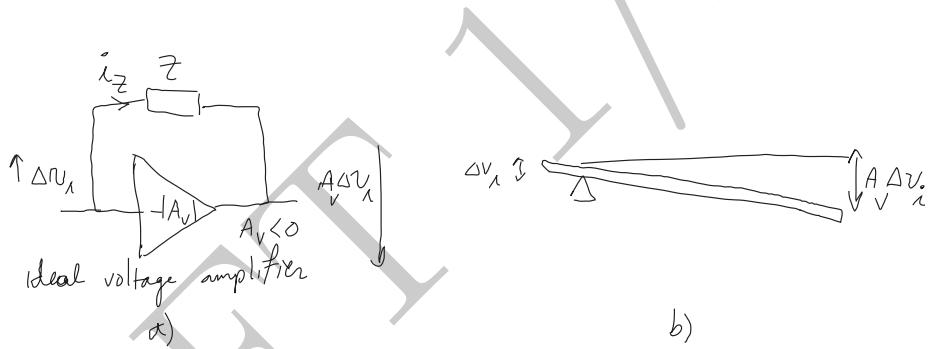


Figure 4.17: a) An ideal inverting amplifier with an impedance between its input and output, b) the seesaw analogy for the behavior of the amplifier, Z splitting into two impedances in series.

the input and output terminals of an ideal voltage amplifier, as shown in Figure 4.17a. Again we could argue that  $Z$  experiences a voltage swing  $1 + |A_v| = 1 - A_v$  greater than an impedance from the input to the ground and hence the equivalent impedance at the input of an *ideal* voltage amplifier with a gain,  $A_v$  and an impedance,  $Z$  connected between its input and output is:

$$Z_1 = \frac{Z}{1 - A_v} \quad (4.29)$$

Another way to see this is noting that for an inverting amplifier when the voltage of the left hand side of  $Z$  goes up, its right hand side voltage is forced down by the amplifier. Using the seesaw analogy of Figure 4.17b, where mechanical displacement is analogous to voltage excursions, we can easily see that there must be a point along  $Z$  where the voltage does not change with input changes,

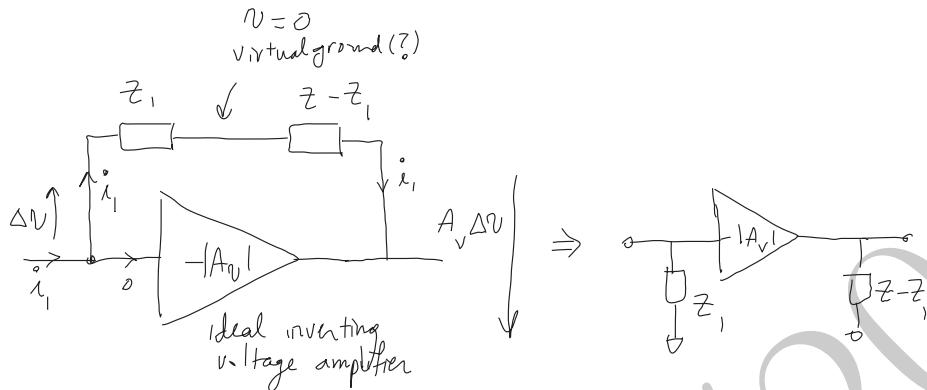


Figure 4.18: a) An ideal inverting amplifier where the feedback impedance,  $Z$ , is split into two impedances in series, b) its equivalent for an *ideal voltage amplifier*.

which corresponds to the pivot point,  $P_1$  in the seesaw analogy. This point has zero electrical potential and can be considered a virtual ground. We can find this point by dividing the impedance  $Z$  into the series combination of two impedances  $Z_1$  and  $Z - Z_1$ , which add up to  $Z$ , as illustrated in Figure 4.18. We notice that the same current flows into both  $Z_1$  and  $Z - Z_1$  hence if the mid-point is at zero volts, we should have

$$\frac{\Delta v}{Z_1} = \frac{-A_v \Delta v}{Z - Z_1}$$

which directly results in (4.29). While we can calculate the second impedance to be  $Z - Z_1 = -A_v Z / (1 - A_v)$ , it is completely inconsequential in this case because it is in parallel with the output impedance of the ideal voltage amplifier, which is zero.

The general behavior of the equivalent input impedance of (4.29) has several important implications in practice. As we saw earlier, a capacitor could be multiplied by a large factor if it is connected across a large inverting gain. This could be an advantage if we need an on-chip capacitor that normally takes up a lot of die area if implemented directly. In that case, it is convenient to use a much smaller capacitor in conjunction with a high-gain inverting amplifier to emulate a much larger capacitor. This situation arises, for instance, when we try to guarantee stability for an amplifier in feedback configuration, as will be discussed in Chapter 6.

On the contrary the same capacitance multiplication (sometimes called ‘Miller Multiplication’) can be very troublesome if we are trying to increase the bandwidth of the stage. As we saw in (4.21) in Example 4.2.5, the resistance seen by the capacitor  $C_\mu$  and hence its time constant is approximately  $|A_v|$  time larger too, which means that there could be a substantial bandwidth limitation due to

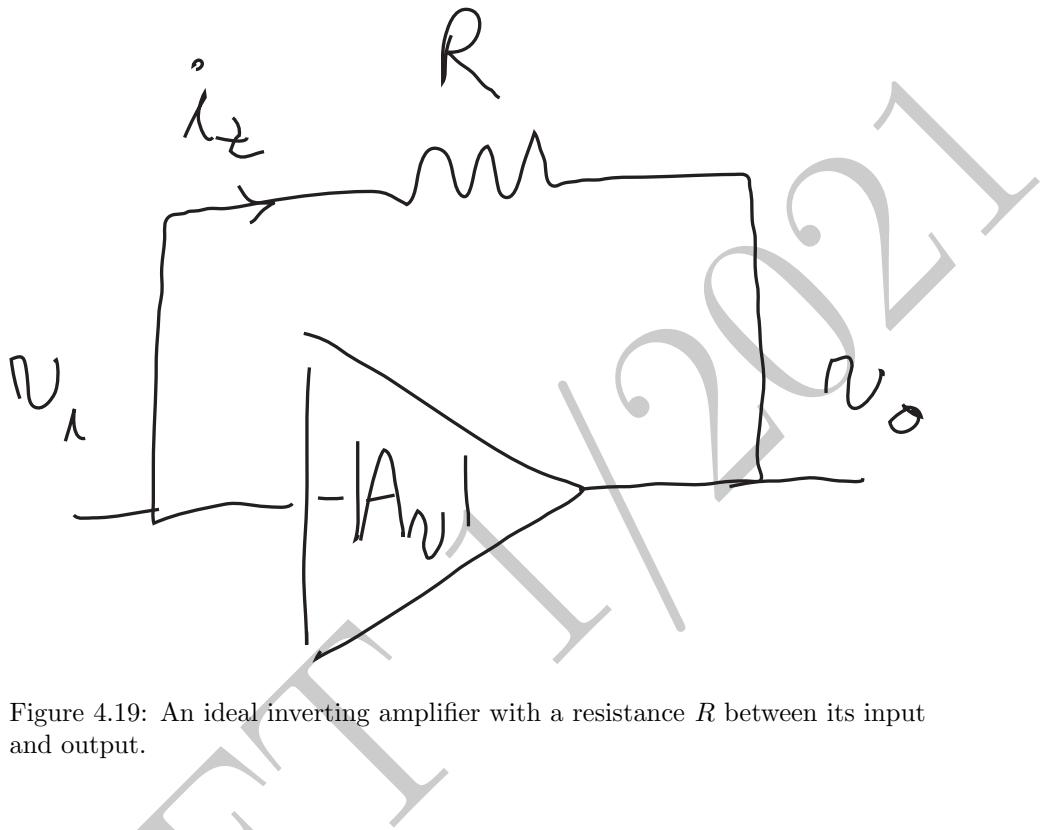


Figure 4.19: An ideal inverting amplifier with a resistance  $R$  between its input and output.

Miller multiplication for large gains. This is in fact often the primary bandwidth limiting factor in common-source/common-emitter amplifying stages, which can be alleviated greatly using the cascode configuration, as we will see later in Example 4.3.4.

Another example of how the result in (4.29) is useful to gain understanding about a circuit is shown in Figure 4.19, where this time a resistor  $R$  is connected across the same ideal inverting voltage amplifier. In this case, according to (4.29) the input impedance is  $R_{in} = R/(1 - A_v)$  which is smaller by the factor  $1 + |A_v|$ . A small input impedance is very useful when we are trying to measure the signal from a signal source that is better approximated by a current source. An example of such a source is a photo-diode used to convert modulated optical signals to electrical domain.

Evaluating (4.29) for a non-inverting gain can also lead to interesting and potentially useful results. First consider the special case when  $A_v = 1$ . In this case, the output voltage tracks the input voltage exactly and the voltage swing across the impedance is simply zero at all times, as depicted in Figure 4.20. When this happens no current flows through  $Z$  and hence it has no bearing on the circuit and can be removed. Therefore, the equivalent input impedance due to  $Z$  is simply an open circuit or infinity. This is consistent with (4.29). This

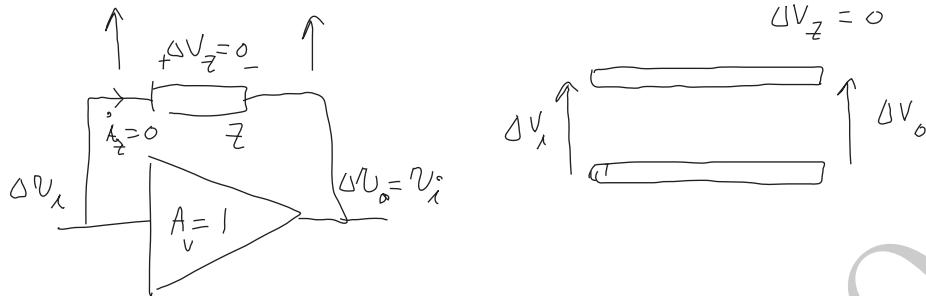


Figure 4.20: An ideal non-inverting amplifier with  $A_v = 1$  and an impedance  $Z$  between its input and output.

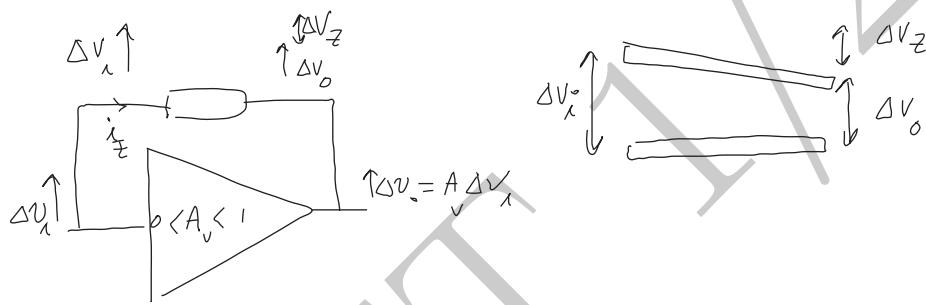


Figure 4.21: An ideal non-inverting amplifier with  $0 < A_v < 1$  and an impedance  $Z$  between its input and output.

can be useful if we are trying to *minimize* the capacitance seen at a given node. All we need to do is to drive the other side of the capacitor in phase and with a gain of one.

Even if the gain is sub-unity but close to one (as is the case in most voltage buffers), we see from (4.29) that the input equivalent impedance is *increased* by a factor of  $1/(1 - A_v)$ . For a capacitor this corresponds to a *reduction* in the input equivalent capacitance. We have already seen an instance of this in Example 4.2.3, where the resistance seen by the capacitor  $C_\pi$  was reduced from the sum of the input and output capacitance by a factor  $1 + g_m R_2$  which is simply  $1/(1 - A_v)$  for the low-frequency voltage gain given by (4.16). A more intuitive way of explaining this is shown in Figure 4.21, where we can see that the output end of the capacitor moves in the same direction as the input but at a slightly smaller rate. Thus, hence the capacitor only experiences a fraction of the input voltage change given by  $(1 - A_v)\Delta v$  and hence the current through it is smaller by the same factor. This is what happens in the source-follower of Example 4.2.3. Note that in this case, attempting to split  $Z$  into two series impedances with a virtual ground in between (Figure 4.17b) results in a negative

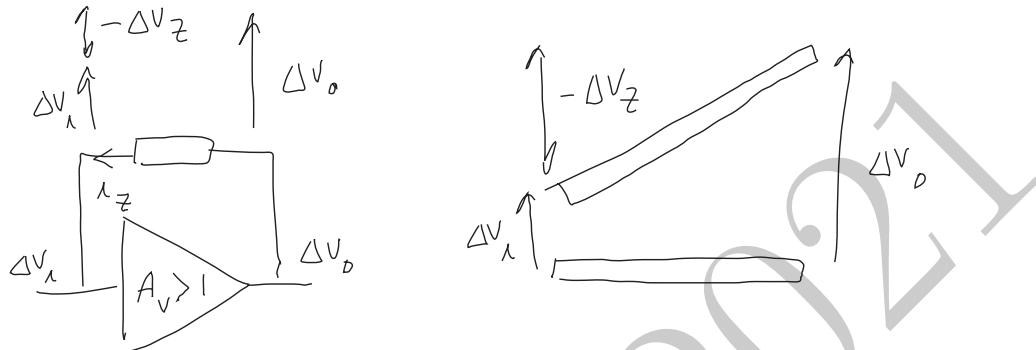


Figure 4.22: An ideal non-inverting amplifier with  $1 < A_v$  and an impedance  $Z$  between its input and output.

impedance  $Z - Z_1$  on the output side. Nonetheless, this is in parallel with the output impedance of the ideal amplifier which is zero.

So far we have looked at the case of an inverting amplifier or a non-inverting amplifier with a gain smaller than or equal to one. It is only natural to ask at this point what happens if we have a non-inverting gain greater than unity, as shown symbolically in Figure 4.22. In this case, the output side of the impedance moves up at a *faster* rate than its input and hence the current flows through it in the *opposite* direction, i.e., an increase in the input voltage results in current being pushed back into the input. This, by definition, corresponds to a *negative* impedance, again in agreement with the result obtained from (4.29). For instance, if  $Z$  is simply a capacitor, the equivalent capacitance at the input will be a *negative capacitor*<sup>17</sup>. This can be a useful element if applied carefully (or cause for a lot of frustration if not).

One practical application of such a negative capacitance is shown in Figure 4.23, where two (usually small) capacitors are cross-connected between the inputs and the outputs. As you can see, each capacitor,  $C$ , is connected across a non-inverting amplifier with a low-frequency gain of  $g_m R_2$ . At the input equivalent capacitance is thus roughly  $(1 - g_m R_2)C$  which is negative if  $g_m R_2 > 1$ . In practical design, this negative input capacitance, whose value is controlled by the choice of  $C$ , can be used to cancel part of the capacitance on the input node to improve the bandwidth. This is an instance of the technique sometimes referred to as *Neutralization*. Though this is a useful trick to be aware of, one should be careful not to introduce too much negative capacitance, as this can lead to excessive peaking and even instability in the limit<sup>18</sup>.

<sup>17</sup>There is nothing improper happening here a negative capacitor is one where its current and the derivative of its voltage have a negative proportionality constant. Also note that a negative capacitor is *not* equivalent to an inductor, as in an inductor the voltage is proportional to the derivative of the current.

<sup>18</sup>One way to see this is by noting that the topology of Figure 4.23 does indeed look similar to a cross coupled oscillator.

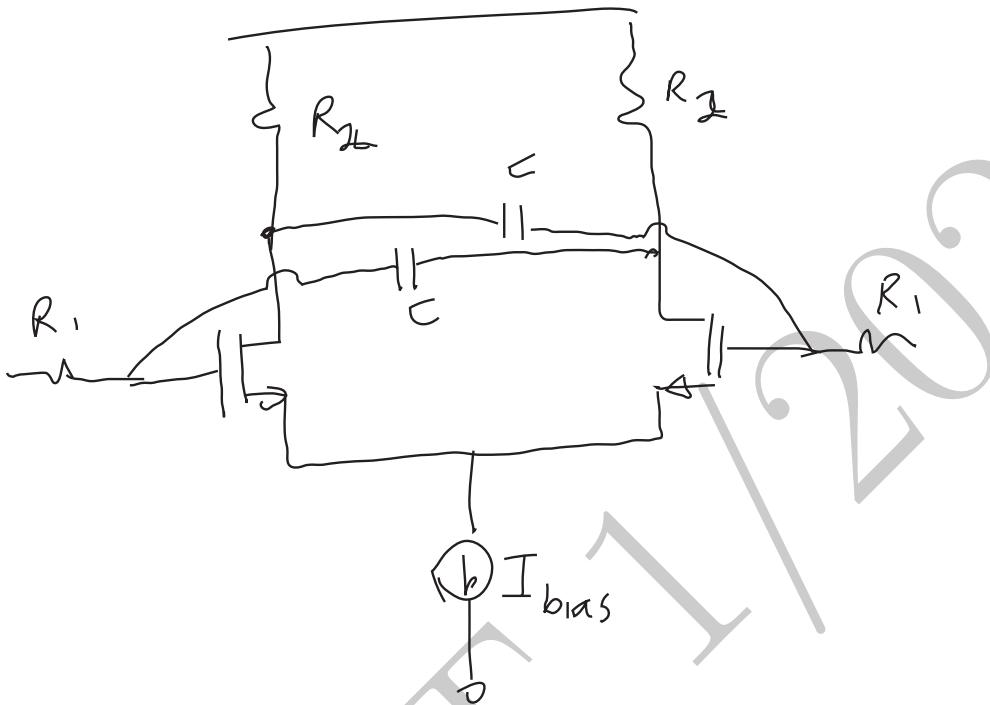


Figure 4.23: Negative capacitance neutralization in a differential pair.

Though quite useful, the result of (4.29) was derived for the ideal voltage amplifier. If the output impedance of the amplifier is not zero, as in the case of the amplifier of Examples 4.2.5 and 4.2.6, shown in Figure 4.10, the virtual ground point in Figure 4.18 can become a function of frequency. Therefore, *it may not be possible* to divide the impedance into the series combination of two impedances in such a way that the mid-point remains at zero potential *for all frequencies*. This is an example where the result of (4.29) becomes approximate, simply because it is applied to a case for which it was not derived. As an example of this, we saw in Example 4.2.6 that the equivalent input impedance has a resistive term of  $r_m \parallel R_2$  in series with  $C_M$ , which (4.29) simply does not predict<sup>19</sup>.

This is not to say that (4.29) cannot provide useful qualitative input to understand the behavior of the circuit, but it is not exactly accurate and in certain cases can produce misleading results<sup>20</sup>. On the other hand the results

<sup>19</sup>Note that  $r_m \parallel R_2 \rightarrow 0$  when  $g_m \rightarrow \infty$  which is the necessary condition for output resistance being zero. This shows the more general nature of the results obtained using the time constant approach of (4.14).

<sup>20</sup>While the Miller approximation is a useful tool to gain insight about electronic circuits and their behaviors, it can be easily misused (abused). The problem arises when the basic underlying principle of this approximation is presented as the almighty Miller “Theorem”

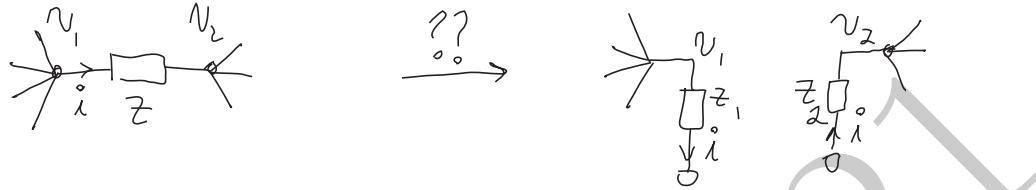


Figure 4.24: a) A single impedance  $Z$  connected between nodes **1** and **2** of a general network b) Two impedances,  $Z_1$  and  $Z_2$  from **1** and **2** to ground.

giving the dangerous sense that it is a provable statement.

The so-called Miller “Theorem” is stated as follows: Consider an impedance  $Z$  connected between nodes **1** and **2** of a general network, as shown in Figure 4.24a. Now *assuming* that this impedance could be replaced by two other impedances,  $Z_1$  and  $Z_2$  to ground, as shown in Figure 4.24b, these two impedances are given by

$$Z_1 = \frac{Z}{1 - A}, \quad Z_2 = \frac{Z}{1 - \frac{1}{A}} \quad (4.30)$$

where  $A$  is the “gain” between node **1** and node **2**, defined as:

$$A \equiv \frac{v_2}{v_1} \quad (4.31)$$

with  $v_1$  and  $v_2$  being the voltages of nodes **1** and **2**, respectively.

The so-called “*proof*” goes as follows: For  $Z_1$  and  $Z_2$  to have the same effect as  $Z$  on the rest of the circuit, the current flowing into  $Z_1$  and out of  $Z_2$  should be equal to the original current flowing through  $Z$ . The current can be expressed as:

$$i = \frac{v_1 - v_2}{Z} = \frac{v_1}{Z_1} = -\frac{v_2}{Z_2} \quad (4.32)$$

which provides two independent equations. Solving them we have

$$Z_1 = \frac{Z}{1 - \frac{v_2}{v_1}}, \quad Z_2 = \frac{Z}{1 - \frac{v_1}{v_2}}$$

which considering (4.31) are equivalent to (4.30) and this “*proves*” the theorem.

As probably self-evident by now, the main problem arises from the *assumption* that there *exist* a pair of frequency independent impedances  $Z_1$  and  $Z_2$  which can emulate the effect of impedance  $Z$  under all conditions and at all frequencies. This initial assumption is where the fallacy begins. As we saw in Figure 4.18 this relies on the existence of the frequency independent virtual ground point, which is not true for non-zero output impedance of the amplifier in general.

To see how this can fail, consider the *first-order* linear circuit of 4.10b, where an input voltage source with a source resistance  $R_1$  drives transconductance  $g_m$  which in turn drives a load resistor  $R_2$ , with a single capacitor  $C_\mu$  connected between the input and the output. We determined in Example 4.2.3 that this system has a LHP pole and a RHP zero. Applying the “Miller theorem” to this circuit will convert it to an equivalent circuit with a capacitance  $C_1 = C_\mu(1 + g_m R_2)$  from the input to ground and a *second* capacitor  $C_2 = C_\mu(1 - 1/g_m R_2)$  between the output node and the ground. This is a preposterous result, as the new “equivalent” circuit is *second order(!)* and furthermore the zero has *disappeared!*

The irony of the situation is that Miller did not present the result as a “theorem”. He used it to explain why a large capacitor was seen at the input for a large inverting gain in an amplifier.

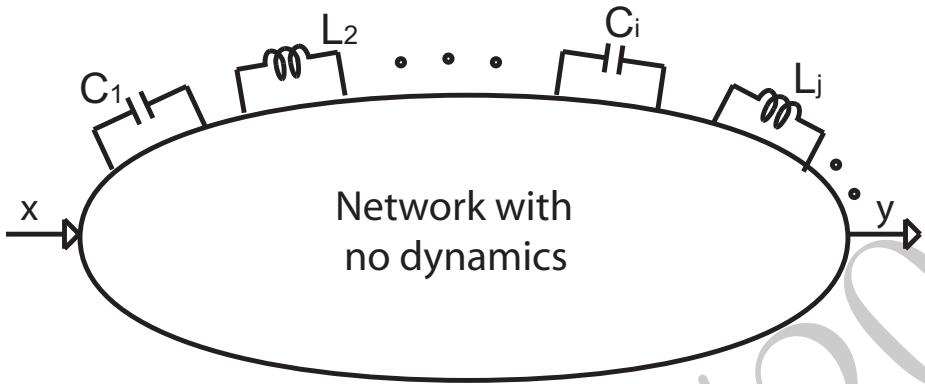


Figure 4.25: An  $N$ -port with all the inductors and capacitors presented at the ports and no energy storing element inside.

obtained from the method of time-constants discussed for a first-order system in section 4.2, which will be generalized to the case of an  $N$ th order system in sections 4.3-4.6, does not suffer from a similar limitation and is exact.

### 4.3 Zero-Value Time Constants

Online YouTube lectures:

[Zero-value time and transfer constants \(ZVT\),  \$b\_1\$  and  \$a\_1\$  term calculation](#)

Having considered a system with one energy storage element, in this section we take the first step toward a generalization of the approach to the case with  $N$  energy storing elements. We will start by trying to determine the first term in the denominator of (4.1), namely,  $b_1$  in a system with  $N$  reactive (energy-storing) elements. We will see that  $b_1$  is *exactly* equal to the sum of the so-called *zero-value time constants* (ZVT) of the network. The zero-value time constants are essentially time-constants of the first-order systems formed by forcing all but one of the reactive elements to be at their zero values, i.e., open-circuited capacitors and shorted-circuited inductors. We will show that the sum of these time constants calculated individually is equal to the coefficient  $b_1$  in the denominator of the transfer function of (4.1).

Any network with  $N$  energy-storing (reactive) elements can be represented as an  $N$ -port with no frequency-dependent elements (e.g., containing only resistors and dependent voltage and current sources) and each reactive element (namely inductors and capacitors) attached to one of the ports<sup>21</sup>, as shown in Figure 4.25. Obviously, the impedances of the capacitor,  $C_i$ , and inductor,  $L_j$ , are

<sup>21</sup>If more than one reactive element is connected to the same pair of terminals, each one of them is assumed to have a port of its own with a separate index.

## ▼ Derivation ▼

$1/C_i s$  and  $L_j s$ , respectively.

The only way for a coefficient  $s$  to occur in a transfer function of a lumped circuit is as a multiplicative factor to a capacitor or an inductor ( $Cs$  or  $Ls$ ). Let us initially limit our discussion to just capacitors and then generalize to include the inductors. In that case, the  $b_1$  coefficient in (4.1) is a linear combination of all the capacitors in the circuit. Then, for instance, the  $b_1$  term cannot contain a term  $C_i C_j$  because such a term must have an  $s^2$  multiplier. Applying the same line of argument, the  $b_2$  coefficient must consist of a linear combination of two-way products of different capacitors ( $C_i C_j$ ), as they are the only ones that can generate an  $s^2$  term<sup>22</sup>. In general the coefficient of the  $s^k$  term must be a linear combination of  $k$ -way products of different capacitors<sup>23</sup>. The same argument can be applied to  $a_k$  coefficients in the numerator and hence we can write the transfer function as

$$H(s) = \frac{a_0 + (\sum_{i=1}^N \alpha_1^i C_i)s + \dots}{1 + (\sum_{i=1}^N \beta_1^i C_i)s + \dots} \quad (4.33)$$

where coefficients  $\beta_1^i$  have units of ohms [ $\Omega$ ] and  $\alpha_1^i$  coefficients have the same units as the transfer function  $H(s)$  times ohms [ $\Omega$ ].

The transfer function of (4.33) is determined independently of the specific value of the capacitor and must therefore be valid for all capacitor values including zero and infinity. The idea behind the derivation of  $a_i$  and  $b_i$  coefficients in general is to choose a set of extreme values (zero and infinity or equivalently open and short) for energy storing elements in such a way that we can isolate one of the  $a$  or  $b$  parameters in question in terms of other parameters we already know and simple low-frequency calculations involving no reactive elements at all. We apply this approach to  $b_1$  in this section and to higher order  $a_i$  and  $b_i$  coefficients in the subsequent ones.

To determine  $b_1$ , let us look at the case when all capacitors, except  $C_i$ , have a value of zero, as depicted in Figure 4.26. This system has only one energy storing element as is thus first order. Its transfer function must consist of first-order polynomials in  $s$  in both the numerator and the denominator. The transfer function of (4.33) with a single  $C_i$  reduces to the following first-order one,

$$H_i(s) = \frac{a_0 + \alpha_1^i C_i s}{1 + \beta_1^i C_i s} \quad (4.34)$$

As a side note, one way to see why the higher order terms in (4.33) are linear combination of products of *different* capacitors with no self product term (e.g.,  $C_i C_i$ ) is by noting that if there were a term  $C_i^2$  in  $a_2$  or  $b_2$  it would lead to a second order transfer function in (4.34) which contradicts the fact that the system in Figure 4.26 has only one energy storing element.

<sup>22</sup>We will see shortly why they cannot be the same capacitor, i.e.,  $(C_i)^2$

<sup>23</sup>There will be more on this subject in Section 4.6.

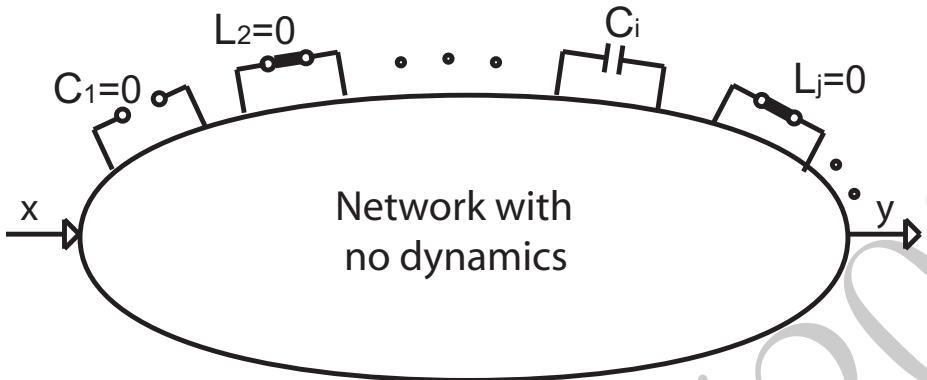


Figure 4.26: An  $N$ -port with all the inductors and capacitors zero valued except  $C_i$ .

We have already discussed such a first-order system in Section 4.2, whose transfer function is given by equation (4.14). The time constant of the first order system of Figure 4.26 with a single capacitor,  $C_i$ , is

$$\tau_i^0 = R_i^0 C_i, \quad (4.35)$$

where  $R_i^0$  is the resistance seen by the capacitor  $C_i$  looking into port  $i$  with all other reactive elements connected to the other ports at their zero value (hence the superscript), namely open-circuited capacitors (and short circuited inductors). Equations (4.14), (4.34), and (4.35) clearly indicate that

$$\beta_1^i = R_i^0 \quad (4.36)$$

which is generally true for the capacitor at any port, since we did not make any assumption valid only for the capacitor  $C_i$ . Hence, the first denominator coefficient in (4.1),  $b_1$ , is simply given by the sum of these *zero-value time constants* (ZVT)<sup>24</sup>,

$$b_1 = \sum_{i=1}^N R_i^0 C_i = \sum_{i=1}^N \tau_i^0 \quad (4.37)$$

where  $\tau_i^0$  coefficients are the ZVT's<sup>25</sup>.

<sup>24</sup>This method is sometimes referred to as the method of open-circuit time constants. This terminology only makes sense when applied to capacitors because a zero-valued capacitor corresponds to an open circuit. Unfortunately, an inductor at its zero value corresponds to a *short circuit* and thus the name becomes misleading. For this reason, we will refer to this method as the method of zero value time constants to maintain its generality.

<sup>25</sup>This result is derived in the literature using an  $n$ -port nodal analysis of the above system and calculations of the co-factors of the circuit determinant (e.g., see the classic work of Thornton, Searle, *et al.* [?] and its later extensions such as the paper by Cochran and Grabel [?]). The physical argument offered here provides better intuition.

The approach remains essentially the same in the case where some of the energy-storing elements are inductors. In that case, the first order terms in the numerator and denominator of (4.1) (i.e.,  $b_1$  and  $a_1$ ) can be written as a linear combination of capacitors and inductors. Let us consider the case of an inductor,  $L_i$ , at port  $i$ , when all the other elements are zero valued. This means open circuited capacitors and short circuited inductors. Again this reduces to the first order system with an inductor that was discussed in Section 4.2 with a time constant similar to the one in (4.15)

$$\tau_i^0 = \frac{L_i}{R_i^0} \quad (4.38)$$

### ▼ Result ▼

Hence, in the general case when both inductors and capacitors are present, the coefficient  $b_1$  in the denominator is exactly given by:

$$b_1 = \sum_{i=1}^N \tau_i^0$$

(4.39)

where the  $\tau_i^0$  terms are zero-value time constants associated with the capacitor or the inductor at port  $i$  given by (4.35) or (4.38), respectively. These time constants are determined by the resistance seen looking into the port, namely  $R_i^0$ , when all other energy storing elements are zero valued (open capacitors and shorted inductors), as illustrated in Figure 4.26

Note that the sum of zero-value time constants in (4.39) is exactly equal to the sum of pole characteristic times ( $-1/p_i$ ) which is also equal to  $b_1$  (see the footnote on page 121). However, it is very important to note that *in general there is no one-to-one correspondence between the individual zero-value time constant,  $\tau_i$ , and pole frequencies,  $p_i$* . (For one thing the individual poles can be complex while the time constants are always real. Also, as we will see later, the number of the poles and the number of time constants are not necessarily the same.)<sup>26</sup>

A simple example demonstrates that there is no one-to-one correspondence between the ZVTs and the poles is given next.

#### **Example 4.3.1** The second order low-pass RC circuit shown in Figure 4.27 is

<sup>26</sup>Sometimes the “Miller theorem” approach discussed in the footnote on page 146 is taken even farther by associating individual poles to the nodes by calculating the “equivalent capacitor” between each node and ground and calculating a pole characteristic time as the  $RC$  time constant of that capacitor and the resistance seen from that node to ground. This is a precarious approach, which only works accurately if all the capacitors have one side already at ground to begin with and their time-constants are not coupled, meaning the resistance seen by one does not change where the other capacitors are open or short (more on why this is the case in the upcoming sections). Also, it is not clear how such an approach applies to inductors. There are few cases where this approach can produce reasonable approximation of the results, but generally it can generate erroneous results with spurious poles that are sometimes even in the right-half plane (e.g. try this with a source follower with gain less than unity for the output node). The methods of this Chapter are far superior to the aforementioned node-pole association and can also tell us quickly when such an approach would produce correct results and when it would not.

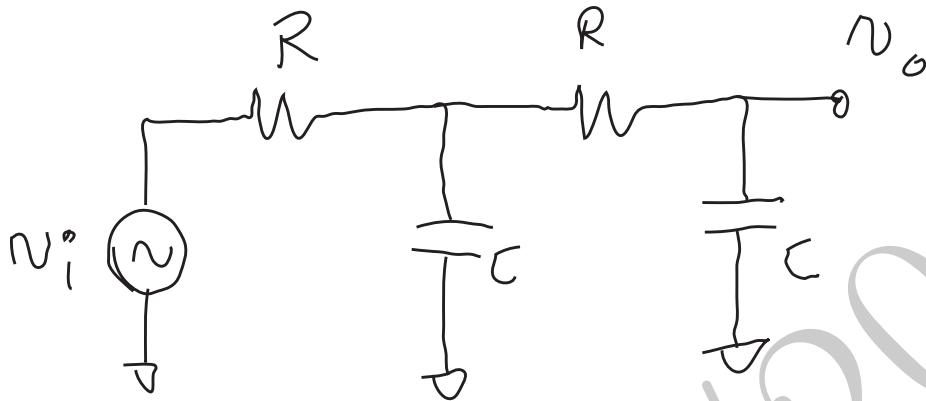


Figure 4.27: A second order low-pass  $RC$  circuit with equal  $R_s$ s and equal  $C$ s.

driven by a voltage source,  $v_i$ . It is easy to see that it has the following time constants associated with the two capacitors<sup>27</sup>

$$\begin{aligned}\tau_1^0 &= RC \\ \tau_2^0 &= 2RC\end{aligned}$$

leading to

$$b_1 = \tau_1^0 + \tau_2^0 = 3RC$$

Using nodal analysis or the method of Sections 4.5 and 4.6, the exact transfer function can be easily determined to be

$$H(s) = \frac{1}{1 + b_1 s + b_2 s^2} = \frac{1}{1 + 3RCs + (RC)^2 s^2}$$

which produces the same  $b_1$  coefficient, as predicted by the ZVTs. However, if one were to incorrectly assume that there is a one-to-one correspondence between the poles and ZVTs, and also that there are no zeros in the transfer function (which is correct in this case), s/he would arrive at

$$H(s) \stackrel{?}{=} \frac{1}{(1 + \tau_1^0 s)(1 + \tau_2^0 s)} = \frac{1}{1 + (\tau_1^0 + \tau_2^0)s + \tau_1^0 \tau_2^0 s^2}$$

which erroneously implies that  $b_2 \stackrel{?}{=} \tau_1^0 \tau_2^0 = 2(RC)^2$ , in contradiction with the directly calculated transfer function that predicts  $b_2 = (RC)^2$ .

The above simple example clearly shows that in general there is no one-to-one correspondence between the ZVTs and the poles because of the coupling between the time constants.

<sup>27</sup>Although in this example the two capacitors are equal in value, we use the index 1 for the one closer to the input and index 2 for the one on the output.

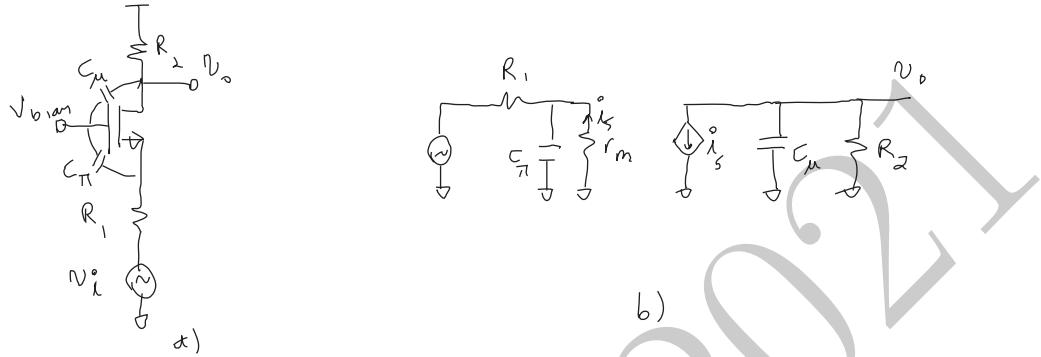


Figure 4.28: a) A common-gate stage driven with a voltage source with a source resistance  $R_1$ , b) small signal model for the common-base stage.

The one-to-one correspondence (i.e.,  $\tau_i^0 \stackrel{?}{=} -1/p_i$ ) is *only* true for those time-constants that are *uncoupled* from the rest, as will be shown in Section 4.6. The time constant,  $\tau_i^0$ , is uncoupled from the rest if the resistance seen looking into port  $i$  remains the same for *each and every* combination of shorting and opening of the remaining ports (or equivalently the energy storing elements). In this case, the term associated with that time constant can be factored out of the denominator and constitutes a single pole at  $p_i = -1/\tau_i^0$ . This is a useful corollary of the time and transfer constants (TTC) method of Section 4.6. Here is an example of decoupled poles:

**Example 4.3.2 Common-Gate** Consider a common gate stage with both  $C_\pi$  and  $C_\mu$  present, as shown in Figure 4.28a with its small-signal model shown in Figure 4.28b. In this case, the time constants associated with the capacitors can be written by inspections to be:

$$\begin{aligned}\tau_\mu &= R_2 C_\mu \\ \tau_\pi &= (R_1 \parallel r_m) C_\pi\end{aligned}$$

Now we can easily see that the resistance seen by each capacitor is unaffected by whether the other one is open and short. Hence the two time constants are uncoupled. Also we notice that either capacitor is short-circuited, the low-frequency transfer function diminishes to zero. Therefore, there are no zeros in the transfer function, and thus it can be written as:

$$H(s) = \frac{H^0}{(1 + \tau_\pi s)(1 + \tau_\mu s)}$$

where  $H^0 = R_2/(R_1 + r_m)$  is the low-frequency gain.

The pole  $p_1 = -1/\tau_\mu$  depends on the value of the load resistance,  $R_2$ , which is multiplied by the relatively small capacitor  $C_\mu$  to produce a typically high

frequency pole. As for  $p_2 = -1/\tau_\pi$  since its resistance to ground is smaller than  $r_m$ , we know that  $p_2$  is on the same order of magnitude as  $\omega_T$  and actually a little bit higher in frequency.

The next subsection deals with application of (4.39) and the ZVTs.

### 4.3.1 Bandwidth Estimation of an Nth-Order System

Online YouTube lectures:

[\*\*Bandwidth estimation using ZVT, high-frequency amplifier design example\*\*](#)

The  $b_1$  coefficient calculated in (4.39) can be used to estimate,  $\omega_h$ , the  $-3dB$  bandwidth of a certain class of low-pass systems<sup>28</sup>. More importantly, it is a powerful design tool allowing the designer to identify the primary source of bandwidth limitation that serves as a guide in making qualitative (e.g., topological) and quantitative (e.g., element values) changes to the circuit.

There are several simplifying assumptions involved in application of the ZVT method to bandwidth estimation. We will eventually introduce the generalized Time and Transfer Constants (TTC) method in Section 4.6, which is capable of determining as many of the  $a_i$  and  $b_i$  coefficients in (4.1) as needed to obtain the desired level of accuracy (up to and including the accuracy obtained by doing full nodal analysis).

In this section, the first simplifying assumption is that there are *no* zeros in the transfer function. Although zeros are present in most circuits<sup>29</sup>, it is a relatively good approximation for many practical systems whose first zero occurs at frequencies much higher than its dominant pole(s) and sometimes close to transistor's cut-off frequency,  $\omega_T$ . Two instances of such systems are the common-source stage of Example 4.2.5 and the source-follower stage discussed in Example 4.2.3 of Section 4.2, where the zero's frequencies are comparable to the cut-off frequency of the transistor itself<sup>30</sup>.

We will spend a fair amount of time in the subsequent sections dealing with the effect of the zeros, what they depend on, when they are generated, and how we can either approximate them quickly or calculate them exactly, but for now let us assume there are no low frequency zeros in the transfer function. A simple test to determine if there are *any* zeros in the transfer function is whether

<sup>28</sup>As we will see later, we can convert a bandpass amplifier, to a low-pass system with the same  $\omega_h$  by setting certain biasing elements such as bypass capacitors, coupling capacitors, and RF chokes to their infinite values (shorted capacitor and open inductor). Then using the method of zero-value time constants we can approximate  $\omega_h$ . A dual process called the method of infinite-value time (IVT) constants discussed in section 4.6.2 can be used to estimate  $\omega_l$ .

<sup>29</sup>We can determine the existence of zeros by applying the method discussed in subsection 4.2.1.

<sup>30</sup>The mere fact that the zero is close to the cut-off frequency of the transistor is not necessarily enough for it to be negligible. As we will see later in Chapter 6, the phase angle of the transfer function around the unity-gain frequency of an amplifier is of utmost importance in its stability when feedback is applied to the amplifier. Having a zero in the close proximity of the unit-gain frequency can have a significant effect on the phase angle of the loop gain.

shorting of any capacitor or opening of any inductor in the circuit results in a non-zero low-frequency transfer function. This would correspond to presence of at least one zero in the transfer function. If there are no dominant zeros in the transfer function, all the frequency dependent terms in the numerator of (4.1) can be ignored and the transfer function can be approximated as

$$H(s) \approx \frac{a_0}{1 + b_1 s + b_2 s^2 + \dots + b_n s^n} \quad (4.40)$$

which is the transfer function of low-pass system with a low-frequency value of  $a_0$ .

Now let us evaluate this transfer function as we increase the frequency from dc to higher frequencies. At dc ( $\omega = 0$ ), the only term in the denominator that matters is the leading 1. As the frequency goes up and starts approaching the  $\omega_h$ , the first term that becomes non-negligible would be  $b_1$ , so in the vicinity of the  $\omega_h$ , (4.40) can be further approximated as a first order system

$$H(s) \approx \frac{a_0}{1 + b_1 s} \quad (4.41)$$

which will imply that  $\omega_h$ , bandwidth of the complete system, will be given by:

$$\omega_h \approx \frac{1}{b_1} = \frac{1}{\sum_{i=1}^N \tau_i^0} \quad (4.42)$$

where  $\tau_i^0$  are the zero value time constants defined by (4.35) and (4.38) for capacitors and inductors, respectively.<sup>31</sup>

Perhaps the most useful aspect of the above approach is its suitability as design tool because it identifies the dominant source of bandwidth limitation. A quick look at the time constants provides a fast and easy way to determine the primary culprit in the circuit. This leads the designer to focus her/his efforts on the first order problem first.

This approximation is conservative and underestimates the bandwidth. To see why, consider (4.40) with only  $b_1$  and  $b_2$  being non-zero as an example. In this case, for  $s = j\omega$  the denominator will be  $1 + b_1 j\omega + b_2 (j\omega)^2 = (1 - b_2 \omega^2) + jb_1 \omega$ . Note that the second order term is making the denominator *smaller* (by making the real part smaller) and hence the transfer function larger<sup>32</sup>, as shown in Figure 4.29. This means that the actual  $\omega_h$  is in fact greater than the one predicted by the first order approximation of the system only considering  $b_1$ .

<sup>31</sup>Intuitively,  $\omega_h$  is the frequency at which the total output amplitude drops by only a factor of  $\sqrt{2}$  with respect to  $a_0$ . Under normal circumstances, at this point the contribution of each one of the energy-storing elements is relatively small and hence (4.42) can be thought of as the sum of their individual contributions to the gain reduction assuming the other ones are not present.

<sup>32</sup>A more accurate approximation for  $\omega_h$  compared to (4.42) can be obtained from this decomposition considering both  $b_1$  and  $b_2$ . It is  $\omega_h \approx 1/\sqrt{b_1^2 - 2b_2}$ .

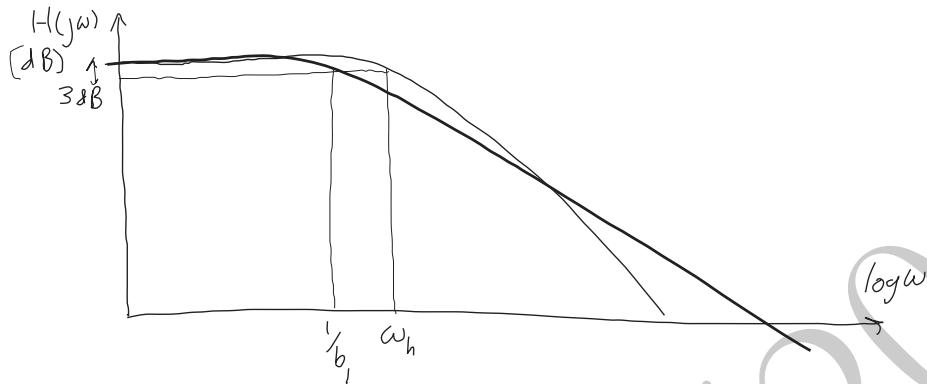


Figure 4.29: The exact transfer function of the  $N$ th order system compared with the first-order approximation based on  $b_1$ .

As mentioned earlier, the coefficient  $b_1$  only provides the sum of the pole characteristic times ( $-1/p_i$ ) with no one-to-one correspondence among  $p_i$ s and  $\tau_i$ s, in general. Therefore, the imaginary parts cancel each other in the  $b_1$  sum since the complex poles of a real system always appear in complex conjugate pairs as discussed earlier. As a result,  $b_1$  does not provide *any* information about the imaginary part of the poles and is completely oblivious to it. This can result in gross underestimation of the bandwidth using (4.42), in cases where the circuit has dominant complex poles which could lead to peaking in the frequency response, as shown in Figure 4.30. An instance of this will be shown in Example 4.3.6. We will also see how we can determine whether or not complex poles are present in Section 4.5.

**Example 4.3.3 Common-Emitter with a Capacitive Load:** Now let us consider the common-emitter stage of Figure 4.31a with both  $C_\pi$  and  $C_\mu$  as well as a capacitive load  $C_L$  between the output (collector) and ground<sup>33</sup>. The equivalent small signal model for this stage is shown in 4.31b.

The dc gain is the product of the gain from the input source,  $v_{in}$ , to the base small-signal voltage,  $v_1$ , (given by the voltage divider ratio between  $R_1$  and  $r_\pi$ ) and the gain from the base to the collector ( $-g_m R_2$ ), i.e.,

$$a_0 = H^0 = \frac{v_o}{v_1} \cdot \frac{v_1}{v_{in}} = -g_m R_2 \cdot \frac{r_\pi}{r_\pi + R_1}$$

Now let us calculate the coefficient  $b_1$ . To do so, we have to calculate three ZVTs associated with each capacitor. In this example, we will use the  $\pi$ ,  $\mu$ , and  $L$  indexes instead of 1, 2, and 3 to provide more insight. To determine the

<sup>33</sup>The case where the load is a parallel combination of a resistor and a capacitor is already subsumed by this example. In that case, we should simply make  $R_2$  be the parallel combination of the collector and load resistors  $R_C \parallel R_L$ . The resistor  $R_2$  can also absorb the output resistance of the transistor,  $r_o$ , into account.

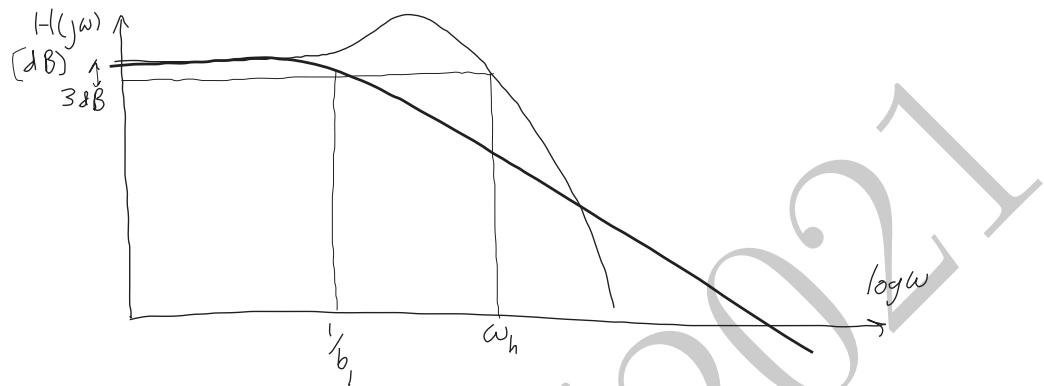


Figure 4.30: The transfer function of the  $N$ th order system with peaking due to complex poles compared with the first-order approximation based on  $b_1$ .

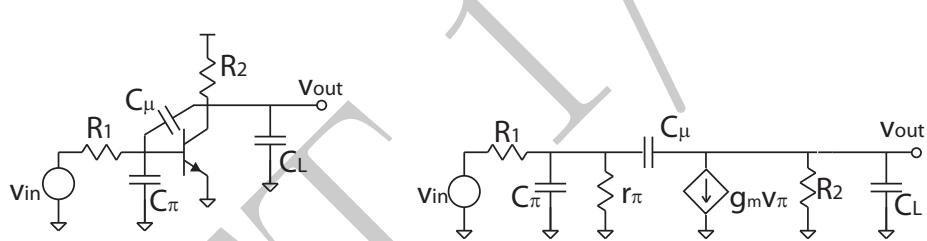


Figure 4.31: a) A common-emitter stage with capacitors  $C_\pi$  and  $C_\mu$  driving a load capacitor,  $C_L$ , b) its small-signal equivalent.

zero-value resistance seen by  $C_\pi$ , we null (short-circuit) the input voltage source and open the capacitors. By inspection, we have

$$R_\pi^0 = R_1 \parallel r_\pi$$

For  $C_\mu$  we have to perform a similar calculation to the one in Example 4.2.5 with a slight modification. Here due to the presence of  $r_\pi$  we have to replace  $R_1$  in (4.20) with  $R_1 \parallel r_\pi$  to obtain:

$$R_\mu^0 = R_1 \parallel r_\pi + R_2 + g_m(R_1 \parallel r_\pi)R_2$$

The calculations for the zero-value resistance seen by  $C_L$  is trivial because nulling the  $v_i$  would set the dependent current source to zero (open circuit) and hence:

$$R_L^0 = R_2$$

Now we can simply apply (4.39) to obtain

$$\begin{aligned}
 b_1 &= \sum_{i=1}^3 \tau_i^0 = \tau_\pi^0 + \tau_\mu^0 + \tau_L^0 \\
 &= C_\pi(R_1 \parallel r_\pi) + C_\mu[R_1 \parallel r_\pi + R_2 + g_m(R_1 \parallel r_\pi)R_2] + C_L R_2 \\
 &= (R_1 \parallel r_\pi)[C_\pi + C_\mu(1 + g_m R_2)] + R_2(C_\mu + C_L) \\
 &= (R_1 \parallel r_\pi)(C_\pi + C_M) + R_2(C_\mu + C_L)
 \end{aligned} \tag{4.43}$$

where  $C_M$  is the Miller capacitance defined by (4.27). This result is consistent with the first-order approximation of the bandwidth obtained from the Miller approximation in Example 4.2.5, showing the multiplication of  $C_\mu$  by the factor  $1 + g_m R_2 = 1 + |A_v|$  similar to (4.27), with the exception of the term  $R_L(C_\mu + C_L)$ . We can estimate  $\omega_h$  as

$$\omega_h \approx \frac{1}{b_1} = \frac{1}{(R_1 \parallel r_\pi)(C_\pi + C_M) + R_2(C_\mu + C_L)} \approx \frac{1}{(R_1 \parallel r_\pi)g_m R_2 C_\mu + R_2 C_L}$$

where  $C_M$  is the Miller capacitance defined by (4.27). The approximation in the last step of the above calculation assumes that the absolute value of the dc gain from base to collector,  $g_m R_2 \gg 1$  and that  $C_M \gg C_\pi$ . Also it is assumed that  $C_L \gg C_\mu$ . All of these assumptions are often true under normal circumstances.

#### ♦ Numerical Example ♦

Let us examine this example numerically for a silicon BJT with the following parameters at the operation point: assuming a collector current of 1mA for now, we have,  $g_m = 40mS$ . Also assume,  $\beta_0 = 100$ ,  $C_{je} = 20fF$ ,  $C_{jc} = 20fF$ ,  $C_{js} = 50fF$ , and  $\tau_F = 2psec$  which corresponds to a  $C_b = g_m \tau_F$  of  $80fF/mA$  at room temperature, leading to  $C_\pi = C_{je} + C_b = 100fF$  and  $C_\mu = C_{jc} = 20fF$ . Now consider an external capacitor on the output  $C_{out} = 150fF$  which together with  $C_{js}$  form  $C_L = C_{out} + C_{js} = 200fF$ . These values correspond to a transistor cut-off frequency,  $f_T \approx 53GHz$ . Assuming  $R_1 = 1k\Omega$  and  $R_2 = 2k\Omega$  in the circuit of Figure 4.31, the low frequency gain is

$$\begin{aligned}
 a_0 &= H^0 = -g_m R_2 \cdot \frac{r_\pi}{r_\pi + R_1} \\
 &= -80 \times 0.714 = -57
 \end{aligned}$$

The time constants are

$$\begin{aligned}
 \tau_\pi^0 &= C_\pi(R_1 \parallel r_\pi) \\
 &= 100fF \cdot (1k\Omega \parallel 2.5k\Omega) \approx 70ps \\
 \tau_\mu^0 &= C_\mu[R_1 \parallel r_\pi + R_2 + g_m(R_1 \parallel r_\pi)R_2] \\
 &= 20fF \cdot (710\Omega + 2k\Omega + 57k\Omega) \approx 1,200ps \\
 \tau_L^0 &= C_L R_2 \\
 &= 200fF \cdot 2k\Omega = 400ps
 \end{aligned}$$

leading to a bandwidth estimate of

$$\omega_h \approx \frac{1}{b_1} = \frac{1}{70ps + 1,200ps + 400ps} \approx 2\pi \cdot 95MHz$$

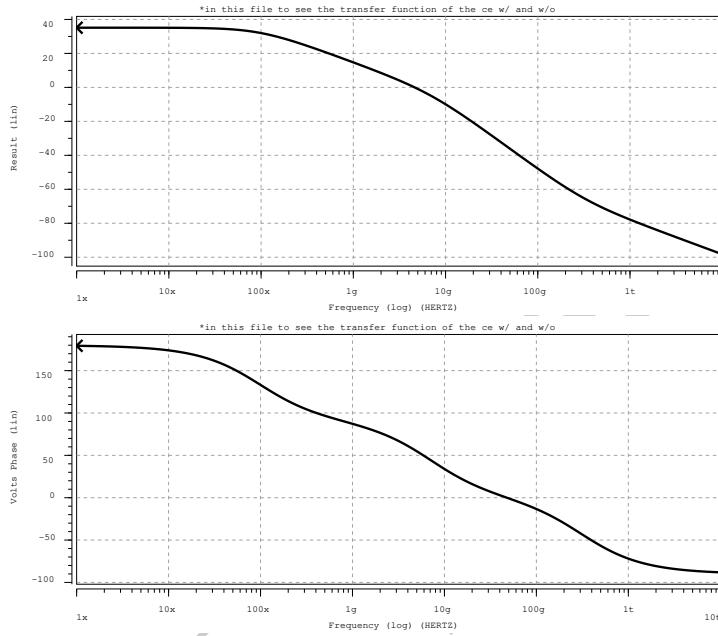


Figure 4.32: The amplitude and phase response of the voltage transfer function of Figure 4.31.

A SPICE simulation predicts a  $-3\text{dB}$  bandwidth of  $f_h = 97\text{MHz}$  in close agreement with the above result and only slightly conservative. The magnitude and the phase of the response obtained from SPICE are shown in Figure 4.32. It is relatively easy to see from Figure 4.32 that there is a second pole around 7GHz where the slope of the amplitude graph changes from  $-20\text{dB/dec}$  to  $-40\text{dB/dec}$  and the phase is approximately  $-135^\circ$ . While not important in the behavior of the circuit in this example, we also notice that there is zero around 300GHz since the amplitude slope goes back to  $-20\text{dB}$ . This is a RHP zero since the phase further decreases due to this zero. The zero is consistent with our prediction of a RHP zero above transistor's  $\omega_T = 2\pi \cdot 53\text{GHz}$  due to the non-zero infinite value gain through  $C_\mu$  discussed in the Example for the common-source.

Although there are three capacitors, they form a capacitive loop and hence there are only two poles. Also, since we have a non-zero infinite-value ( $H^\mu$ ) only for  $C_\mu$ , with a different polarity from  $H^0$ , we should expect RHP zero. This is in agreement with the above simulation results.

As can be seen from the last example, the time constant,  $\tau_\mu^0$ , is the largest source of bandwidth limitation followed by load time constant,  $\tau_L^0$ . This is remarkable since  $C_\mu$  is by far the smallest capacitor among the three considered here. It should be clear that this is due to the *Miller Effect* discussed earlier.

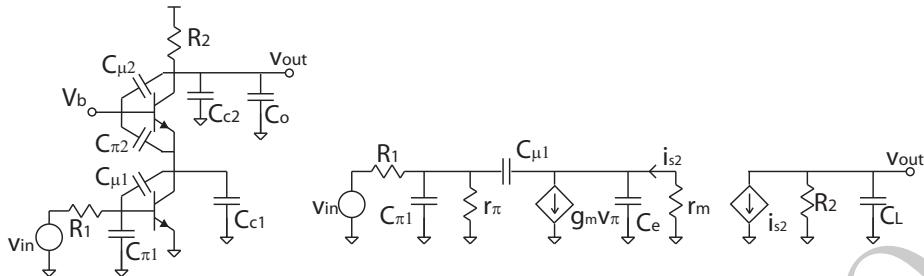


Figure 4.33: a) A cascode stage driving a load capacitor,  $C_L$ , b) its small-signal equivalent.

Namely, the righthand side of  $C_{\mu}$  moves in the opposite direction by a large factor (80 in this case) for a small change in the input, thus the capacitor appears roughly 1+80 times larger at the input.

If we were to improve something in the above circuit, it must start with dealing with large value of  $\tau_{\mu}^0$  due to Miller multiplication. There are several ways to deal with this and we will discuss some of the other approaches later, but the most common way is to reduce the voltage swing across  $C_{\mu}$  by making the transistor drive the emitter of a second transistor connected in a common-base configuration which in turn drives the load resistor  $R_2$ . The main point here is that the impedance looking into the emitter (or source) of a transistor that serves as the collector load for the first transistor is small (on the order of  $r_m$ ) and hence the gain from its base to its collector is on the order of  $-g_{m1}r_{m2}$  which is very large and is typically close to  $-1$ . This reduces the Miller multiplication factor to roughly 2. The second transistor is connected in common-base configuration which has no capacitor connected between its input (emitter) and its output (collector), and hence no significant Miller multiplication. The cascode stage is further studied in the following example.

**Example 4.3.4 Cascode Stage:** The cascode stage is illustrated in Figure 4.33a with its small signal equivalent circuit in Figure 4.33b. Capacitors  $C_{c1} = C_{js}$  and  $C_{\pi 2}$  are in parallel, so are  $C_{c2} = C_{js}$ ,  $C_{\mu 2}$ , and  $C_o$  and hence we define,  $C_e = C_{c1} + C_{\pi 2}$ , and  $C_L = C_{c2} + C_{\mu 2} + C_o$  and deal with four capacitors from this point on.

Let us calculate the time constants again and compare them with the case of common-emitter in the previous example:

$$\tau_{\pi}^0 = C_{\pi}(R_1 \parallel r_{\pi})$$

$$\tau_{\mu}^0 = C_{\mu}[R_1 \parallel r_{\pi} + \alpha r_m + g_m(R_1 \parallel r_{\pi})\alpha r_m] \approx C_{\mu}[2(R_1 \parallel r_{\pi}) + r_m]$$

$$\tau_e^0 = C_e \alpha r_m$$

$$\tau_L^0 = C_L R_2$$

As you can see  $\tau_{\pi}^0$  has not changed, while  $\tau_{\mu}^0$  which was the major bottleneck in

design has been reduced substantially (roughly by a factor of  $g_m R_2 / 2$ ). The load time-constant,  $\tau_L$ , has only marginally increased due to the addition of  $C_{\mu 2}$ , and the new time constant  $\tau_e^0$  is comparable to the reciprocal of transistor's cut-off frequency,  $1/\omega_T$ . Now the relative significance of  $\tau_e^0$  and  $\tau_L$  depends on the values of  $R_1$ ,  $R_2$ ,  $I_C$ , and  $\beta_0$ .

### ♦ Numerical Example ♦

Numerically, we have  $C_e = C_{c1} + C_{\pi 2} = 150 fF$  and  $C_L = C_{c2} + C_{\mu 2} + C_o = 220 fF$ . Let us calculate the time constants again and compare them with the case of common-emitter in the previous example:

$$\begin{aligned}\tau_\pi^0 &= C_\pi (R_1 \parallel r_{\pi 1}) \\ &= 100 fF \cdot (1 k\Omega \parallel 2.5 k\Omega) \approx 70 ps \\ \tau_\mu^0 &= C_\mu [R_1 \parallel r_{\pi 1} + \alpha r_{m2} + g_m (R_1 \parallel r_{\pi 1}) \alpha r_{m2}] \\ &\approx 20 fF \cdot (710 \Omega + 25 \Omega + 710 \Omega) \approx 28 ps \\ \tau_e^0 &= C_e \alpha r_{m2} \\ &\approx 150 fF \cdot 25 \Omega \approx 4 ps \\ \tau_L^0 &= C_L R_2 \\ &= 220 fF \cdot 2 k\Omega = 440 ps\end{aligned}$$

where we can clearly see that  $\tau_\mu^0$  is not significant any more. We also see that in this case  $\tau_L^0$  is now clearly the limiting factor. Estimating  $\omega_h$  based on the ZVTs, we have

$$\omega_h \approx \frac{1}{b_1} = \frac{1}{70 ps + 28 ps + 3.8 ps + 440 ps} \approx 2\pi \cdot 294 MHz$$

where SPICE simulations indicate a -3dB frequency of  $f_h = 337 MHz$ . The magnitude and the phase of the response obtained from SPICE are shown in Figure 4.34. From the amplitude and the phase plot we can see that there are three poles: one roughly around 400MHz, the second around 1.5GHz, and a third one in the 40GHz neighborhood. The change in the slope of the magnitude plot in conjunction with the further reduction in phase around 40GHz indicates a RHP zero there. This behavior is expected as there are four separate capacitors in the circuit ( $C_{\pi 1}$ ,  $C_{\mu 1}$ ,  $C_{e2}$ , and  $C_L$ ), however, capacitors  $C_{\pi 1}$ ,  $C_{\mu 1}$ , and  $C_{e2}$  form a capacitive loop and hence we only can define three independent initial conditions and thus have three poles. The only capacitor shorting of which results in a non-zero transfer function is  $C_{\mu 1}$  and hence we expect a zero which is RHP since  $H^0$  and  $H^\mu$  have opposite polarities.

We saw in the previous example that the load capacitance and its associated ZVT became the bottleneck. Next, we will see if adding a common-emitter buffer amplifier to drive the load capacitor is going to help.

**Example 4.3.5 Cascode Stage with Output Buffer:** The cascode amplifier with an output buffer is shown in Figure 4.35a with its small signal equivalent circuit in Figure 4.35b. This time the parallel capacitors,  $C_{c2}$ ,  $C_{\mu 2}$ , and  $C_{\mu 3}$  are

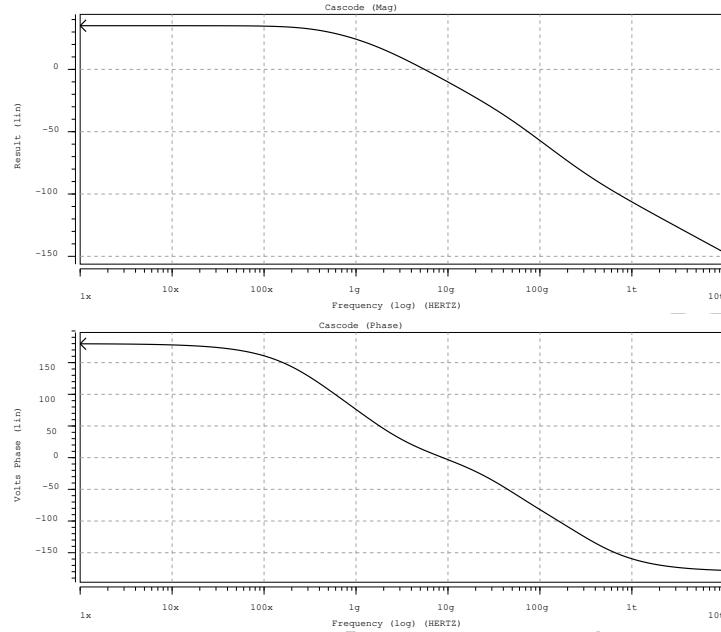


Figure 4.34: The amplitude and phase response of the voltage transfer function of Figure 4.33.

lumped together into  $C_c = C_{c2} + C_{\mu 2} + C_{\mu 3}$ , while we still have  $C_e = C_{c1} + C_{\pi 2}$ . We note that there are six separate capacitors and two independent capacitive loops, one formed by  $C_{\pi 1}$ ,  $C_{\mu 1}$ , and  $C_e$ , and a second capacitive loop formed by  $C_c$ ,  $C_{\pi 3}$ , and  $C_L$ . Hence we expect to have four poles, while we have six ZVT time constants.

In evaluating the time constants, we notice that the ZVTs associated with  $C_{\pi 1}$ ,  $C_{\mu 1}$ ,  $C_e$ , are the same as the previous example, and hence we will not recalculate them. Now the remaining time constants are

$$\begin{aligned}\tau_c^0 &= C_c R_2 \\ \tau_{\pi 3}^0 &= C_{\pi 3} \alpha r_{m3} \\ \tau_L^0 &= C_L (\alpha r_{m3} + \frac{R_2}{1 + \beta_0}) \approx C_L \frac{r_{\pi 3} + R_2}{\beta_0}\end{aligned}$$

The capacitance across  $R_2$  is reduced and hence  $\tau_c^0$  is not as large as  $\tau_L^0$  in the previous example. These new time-constants must be evaluated.

We have  $C_c = C_{c2} + C_{\mu 2} + C_{\mu 3} = 90 fF$ . Let us calculate the the time constants again and compare them with the case of common-emitter in the previous

♦ Numerical Example ♦

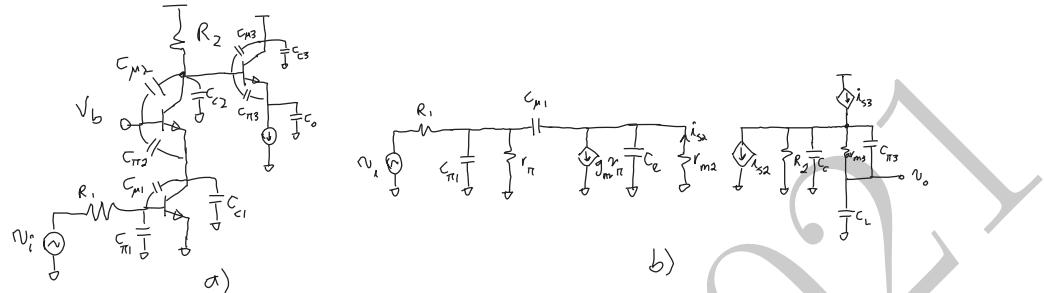


Figure 4.35: a) A cascode stage with a common-collector buffer driving the load capacitor,  $C_L$ , b) its small-signal equivalent.

example:

$$\begin{aligned}\tau_c^0 &= C_c R_2 \\ &= 90fF \cdot 2k\Omega = 180ps \\ \tau_{\pi 3}^0 &= C_{\pi 3} \alpha r_{m3} \\ &= 100fF \cdot 25\Omega = 2.5ps \\ \tau_L^0 &= C_L \left( \alpha r_{m3} + \frac{R_2}{1 + \beta_0} \right) \\ &= 150fF \cdot (25\Omega + 20\Omega) \approx 7ps\end{aligned}$$

Using all six ZVTs we obtain

$$\omega_h \approx \frac{1}{b_1} = \frac{1}{70ps + 28ps + 4ps + 180ps + 2.5ps + 7ps} \approx 2\pi \cdot 546MHz$$

where SPICE simulations indicate a -3dB frequency of  $f_h = 717MHz$ , which corresponds to a factor of two improvement over the previous example. Again we notice the conservative nature of the ZVT estimate. The magnitude and the phase of the response obtained from SPICE are shown in Figure 4.36. As mentioned earlier we expect four poles. We also expect a RHP zero due to  $C_{\mu 1}$  as before. This time we also have a LHP zero due to  $C_{\pi 3}$  shorting of which also results in a non-zero transfer function, which has the same polarity as that when it is open and hence we get a LHP zero. From Figure 4.36 there are two poles roughly around 1GHz, one around 15GHz, and a LHP pole-zero pair around 50GHz. The RHP zero is around 300GHz.

One thing worth mentioning here is that in the sequence of the above three examples, we never removed a single capacitor, and rather added new ones every time we introduced a new transistor. Nonetheless, the topological changes resulted in the capacitors having a smaller impact on the bandwidth (i.e., seeing a smaller resistor and have a larger time-constant). This is an example of how

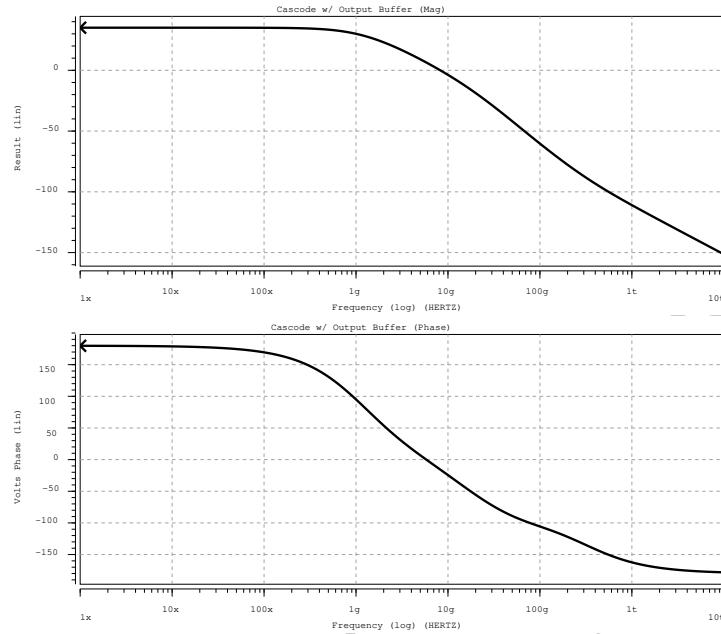


Figure 4.36: The amplitude and phase response of the voltage transfer function of the cascode stage with a common-collector buffer, shown in Figure 4.35.

design-oriented analysis tools provide us with a way to introduce qualitative changes in the circuit beyond that provided through the optimization of the parameters for a given stage

In the examples, we looked at so far, equation (4.42) was rather successful in providing us with a reasonable estimate of the bandwidth and more importantly with a means to identify the primary source of bandwidth limitation in a circuit. The following examples as well as Example 4.4.1 in the next section, provide us some insight into when the bandwidth estimation aspect of ZVT method fails or needs to be modified. This is when the more general tools of the subsequent sections will come handy.

Online YouTube lectures:

#### [When ZVT bandwidth estimation fails](#)

**Example 4.3.6 Source-Follower with Capacitive Load:** Consider the source-follower stage of Figure 4.37, with three capacitors,  $C_\pi$ ,  $C_\mu$ , and  $C_L$ . Noting that the three capacitors form a capacitive loop, we expect two poles. Also noting that  $C_\pi$  is the only capacitor shorting of which produces a non-zero transfer function, and that when shorted the polarity of the transfer function does not change, we expect a LHP zero.

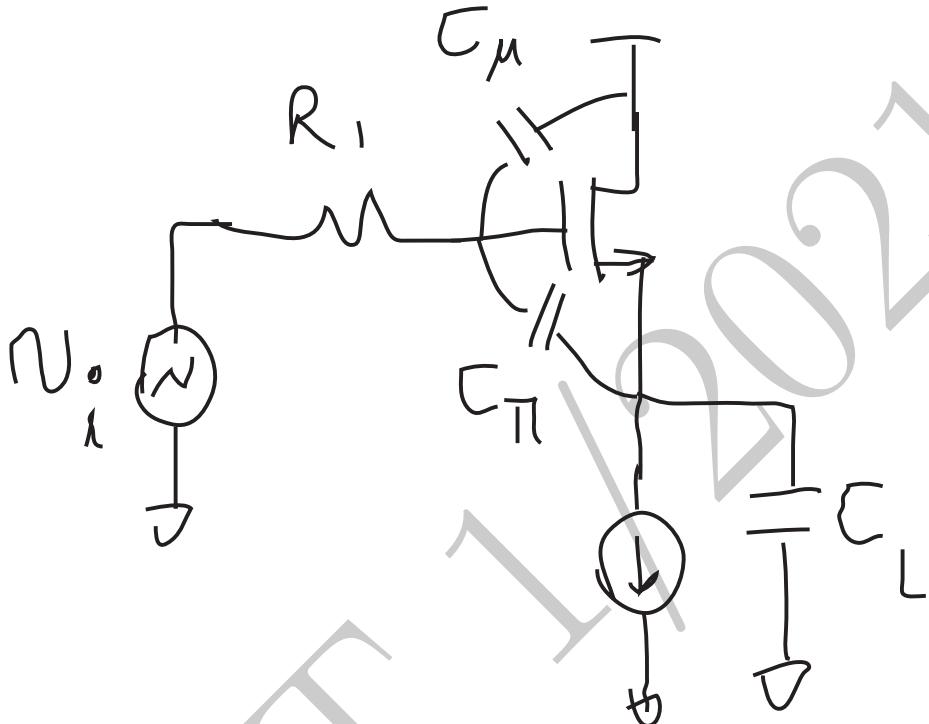


Figure 4.37: A source follower stage with a capacitive load.

The ZVTs are:

$$\tau_\pi^0 = C_\pi r_m$$

$$\tau_\mu^0 = C_\mu R_1$$

$$\tau_L^0 = C_L r_m$$

which we can use to estimate the bandwidth, according to (4.42), we have

$$\omega_h \approx \frac{1}{b_1} = \frac{1}{\tau_\pi^0 + \tau_\mu^0 + \tau_L^0} = \frac{1}{r_m C_\pi + R_1 C_\mu + r_m C_L}$$

#### ♦ Numerical Example ♦

Now let us consider this for a set of numerical values. Both  $C_\pi$  and  $C_L$  are equal<sup>34</sup> to  $50fF$  and  $C_\mu = 10fF$ . Also assuming  $g_m = 20mS$  and  $R_1 = 100\Omega$ , we have,

$$\omega_h \approx \frac{1}{b_1} = \frac{1}{2.5ps + 1ps + 2.5ps} = 2\pi \cdot 26.5GHz$$

<sup>34</sup>This is a reasonable assumption since in many cases the output of the source-follower drives the input of a common-source of similar size, which will have an input capacitance of  $C_\pi$ .

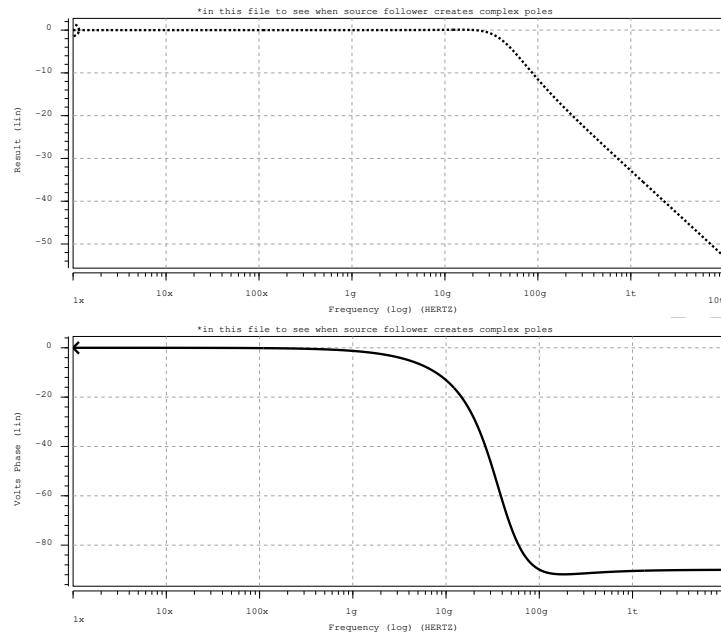


Figure 4.38: The amplitude and phase response of the voltage transfer function of the source-follower driving a capacitive load, shown in Figure 4.37.

However, a SPICE simulations indicate a -3dB frequency of  $f_h = 45\text{GHz}$ . Although, the approximation does remain conservative, it is perhaps too conservative to be useful. The magnitude and the phase of the response obtained from SPICE are shown in Figure 4.38, which might not appear out of the ordinary at first.

Nonetheless, we must explain the large error in our bandwidth estimate. A closer look at the graphs reveal a very minor peaking (about 0.25dB) and a faster than usual drop in the phase (around 80°/dec). These are both signs of complex poles<sup>35</sup>. Since we expect to have only two poles, it means that they must come as a complex conjugate pair. Because  $b_1$  only provides us with the sum of the pole characteristic times (i.e.,  $-1/p_i$ ), all information about the imaginary part of any complex conjugate pair of poles will be lost because they cancel each other.

A corollary of this example is that if there are complex conjugate poles in the transfer function, ZVT may produce excessively conservative estimates for  $\omega_h$ . It also provides no information about the existence and the extent of any peaking in the frequency response. We will see in Section 4.5 how we can detect

<sup>35</sup>One might suspect that the zero may be the culprit, but since it is LHP, it can only slow down the phase drop rate and hence cannot explain the behavior seen.

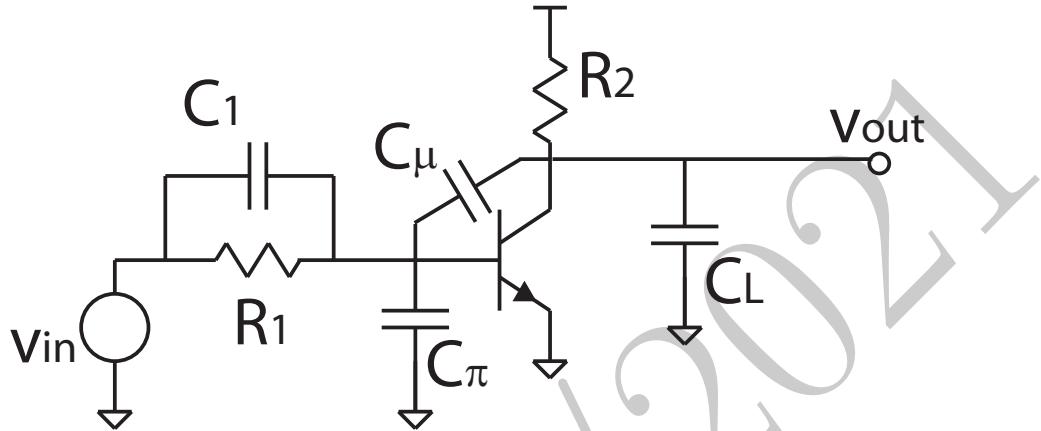


Figure 4.39: a) A common-emitter stage with a capacitor  $C_1$  in parallel with the input resistance  $R_1$ .

these situations and provide a more accurate estimate “for the bandwidth and peaking.

**Example 4.3.7 Common Emitter with Parallel RC in Series with the Input:** Let us consider the common emitter of Example 4.3.3 where a capacitor  $C_1$  is introduced in parallel with  $R_1$  at the input, as shown in Figure 4.39.

The time constants calculated in Example 4.3.3 remain the same. Only a new time constant,  $\tau_1^0$ , associated with  $C_1$  will appear in  $b_1$ , which is easily calculated to be

$$\tau_1^0 = C_1(R_1 \parallel r_\pi)$$

Hence the new bandwidth estimate based on (4.39) is

$$\omega_h \approx \frac{1}{b_1} = \frac{1}{(R_1 \parallel r_\pi)(C_\pi + C_1 + C_M) + R_2(C_\mu + C_L)}$$

which is smaller than the predicted  $\omega_h$  when  $C_1$  is not there. Now let us look at a numerical case.

If  $C_1 = 4.3\text{pF}$  for the input capacitor<sup>36</sup> and all other values are the same as those in Example 4.3.3, we have  $\tau_1^0 \approx 3.07\text{ns}$  and the bandwidth estimate according to (4.39) will be

$$\omega_h \approx \frac{1}{b_1} = \frac{1}{70\text{ps} + 1,200\text{ps} + 400\text{ps} + 3,070\text{ps}} \approx 2\pi \cdot 34\text{MHz}$$

However, a SPICE simulation predicts a -3dB bandwidth of  $f_h = 482\text{MHz}$  which is more than an order of magnitude higher! The magnitude and the phase of the response obtained from SPICE are shown in Figure 4.40.

<sup>36</sup>The reason for the choice of this value becomes apparent later.

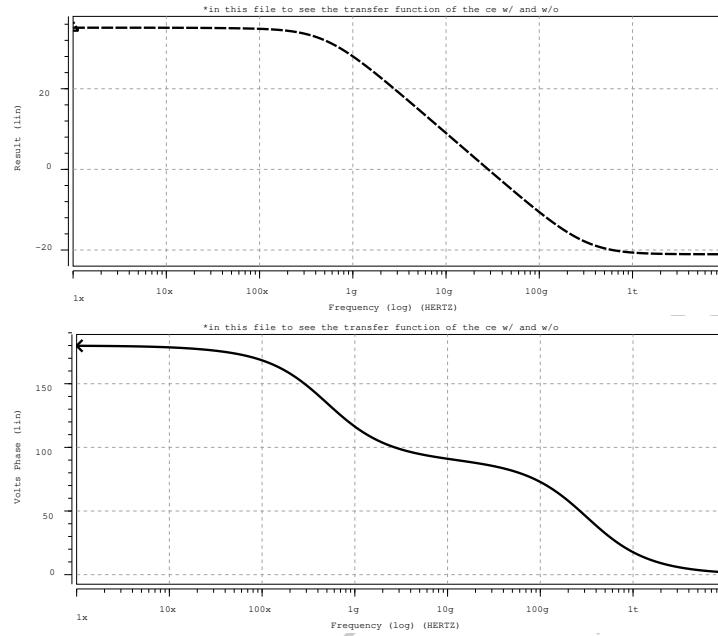


Figure 4.40: The amplitude and phase response of the voltage transfer function of Figure 4.39.

*There is clearly somewhat of a disconnect here. As you may have guessed,  $C_1$  introduces a LHP zero that can be adjusted to coincide with the first pole of the transfer function effectively canceling it, as will be discussed in greater details later in Examples 4.4.1 and 4.4.2.*

In the last example, although (4.39) is still providing a conservative value, it is too far off to be of much use<sup>37</sup>. The basic premise for the approximation in (4.39) was the absence of any zero close to or below  $\omega_h$ . Once that assumption is violated, (4.39) does not provide as much useful information.

One important point that must be emphasized here is that (4.39) predicts the value of  $b_1$  exactly and does not make any claims about its relationship to the bandwidth. In subsection 4.3.1, we assumed the system was low-pass with a zero and applied the method to it. In the case of a bandpass system, (4.39) still produces the correct  $b_1$ , however, in such a case,  $b_1$  does not determine  $\omega_h$ . We will discuss this in subsection 4.6.2 where we talk about the method of infinite-value time (IVT) constants. This is not to say that ZVT cannot predict  $\omega_h$  in a bandpass system. To be able to use the ZVT to determine the high-frequency

<sup>37</sup>On a lighter note, one could always be technically right if s/he predicts a minimum bandwidth of zero, which is technically correct but absolutely useless.

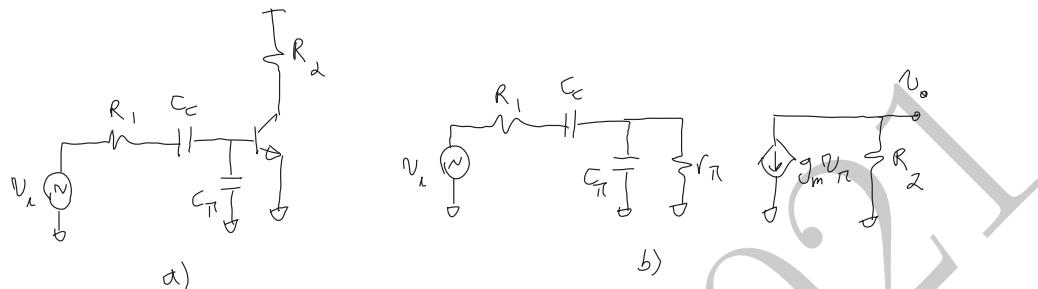


Figure 4.41: a) A common-emitter stage with a coupling capacitor,  $C_c$  at the input, b) its small-signal equivalent.

bandwidth of this circuit,  $\omega_h$ , we must modify *the circuit itself* in such a way that  $b_1$  corresponds to the  $\omega_h$ . This is done by converting the circuit with a bandpass transfer function such as in Figure 4.2 to its low-pass equivalent by removing all the *inverse poles* and *zeros* in the transfer function of (4.4).

The most straightforward way to deal with this is to set all the reactive elements that limit the low-frequency bandwidth ( $\omega_l$ ) to their infinite values (shorted capacitors and opened inductors) and apply the ZVT to the resultant lowpass circuit. The easiest way to determine which reactive elements affect the low frequency response, one can compare the transfer function with that element at its zero and its infinite values. Only those reactive elements whose infinite values result in a larger transfer function at should be infinite valued.

There is similar procedure for determining  $\omega_l$  using the method of infinite value time-constants (IVT) that will be discussed in subsection 4.6.2. We will also develop a generalized method in section 4.6 that subsumes all of these approaches.

Now let us apply this approach to a simple bandpass amplifier.

**Example 4.3.8 AC-Coupled Common-Emitter:** Consider the common-emitter stage of Figure 4.41a, where there is a coupling capacitor,  $C_c$ , between the input source resistance and the input of the stage. We ignore  $C_\mu$  in this example and focus on  $C_\pi$  alone<sup>38</sup>, which is much smaller than  $C_c$  in a typical design to provide a flat gain in the mid-band. The biasing details are not shown. Without doing any analysis, we can easily tell that the amplifier of Figure 4.41b is a bandpass amplifier, since no matter how large we make  $C_c$ , it does not pass dc, hence  $a_0 = 0$ . Also we can see that at very large frequencies  $C_\pi$  short-circuits the input of the amplifier to ground and diminishes the gain. The gain of this amplifier vs. frequency is shown in Figure 4.42.

If we do not modify the circuit to remove the IVT elements, the coefficient  $b_1$  will be dominated by  $C_c$  which determines the first pole responsible for flattening

<sup>38</sup>While both  $C_\mu$  and  $C_\pi$  can be taken into account here, we do not gain any additional insights by including both in this case.

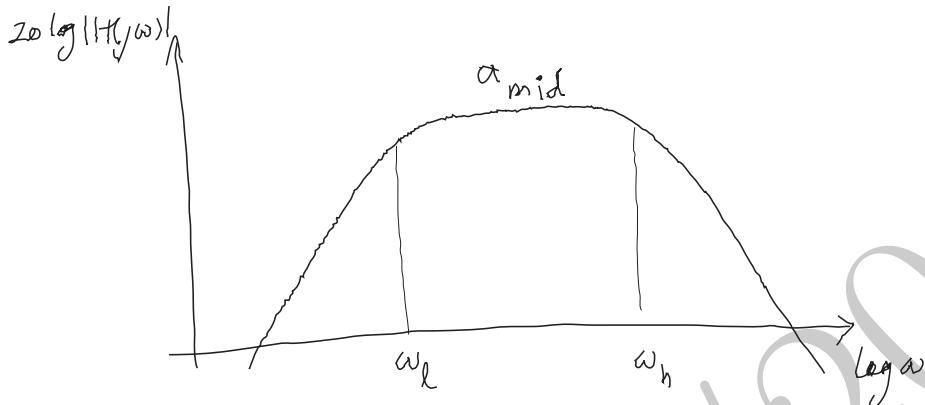


Figure 4.42: The transfer function of the common-emitter stage of Figure 4.41.

of the response at  $\omega_l$ . We know that we can always calculate  $b_1$  exactly, so we must modify the circuit in such a way to make  $b_1$  be a good approximation of  $\omega_h$ .

We notice that the coupling capacitor is intended to behave as an approximation of a short-circuit and the low cut-off frequency is simply a manifestation of its failure to do so below certain frequencies. To eliminate the low frequency dynamics and estimate  $\omega_h$  using zero-value time-constant calculations, we should simply replace  $C_c$  with its ideal “role model,” namely a short circuit.

Replacing  $C_c$  with a short, the circuit reduces to that shown in Figure 4.43 with a single capacitor,  $C_\pi$ . The response of this system can be easily calculated noting that the new  $b_1$  coefficient is

$$b'_1 = \tau_\pi^0 = R_\pi^0 C_\pi = (r_\pi \parallel R_1) C_\pi$$

which determines the  $\omega_h$ .

To summarize, we notice that we can use the ZVT method to determine the  $\omega_h$  of band-pass amplifiers by converting it to an equivalent low-pass amplifier as shown in Figure 4.43. This is done by setting the coupling and bypass capacitors to their ideal value, i.e., a short-circuit and evaluating the zero-value time-constants appropriately. We will revisit example 4.3.8 in subsection 4.6.2 when we discuss the infinite value time-constants.

## 4.4 Zero-Value Time Constants and Transfer Constants for Zeros

INTERMEDIATE TOPIC

Online YouTube lectures:

[Taking zeros into account in ZVT bandwidth estimate](#)

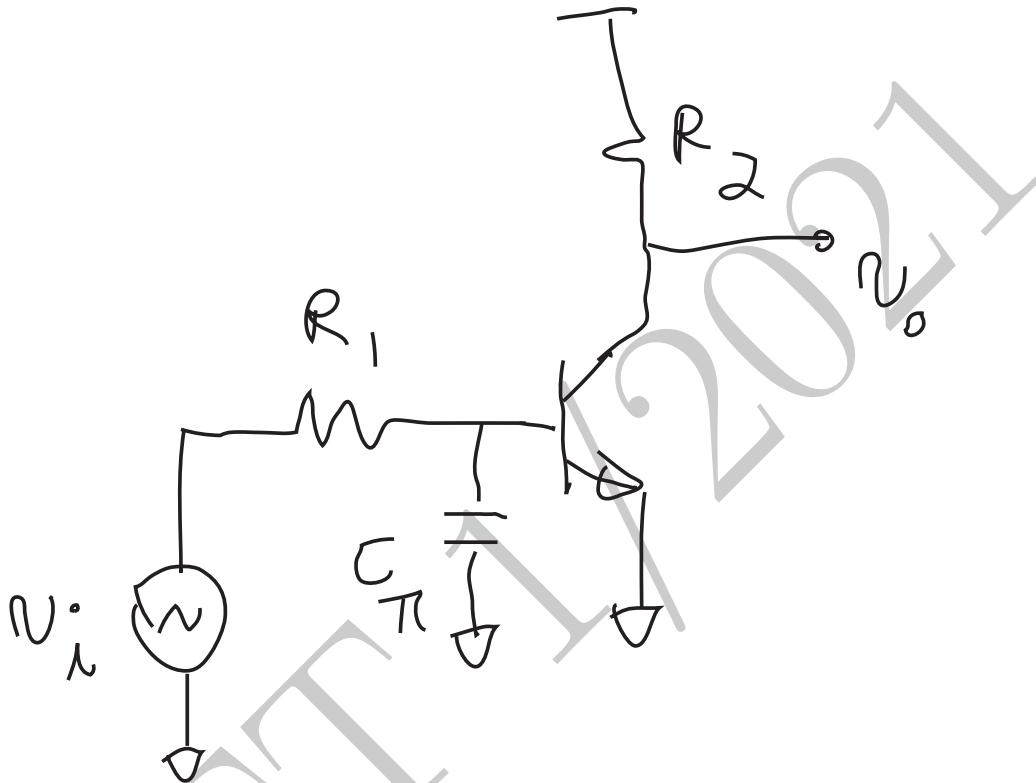


Figure 4.43: The high-frequency equivalent for the common-emitter stage of Figure 4.41.

Now we will determine coefficient  $a_1$  which can be used to approximate the effect of the zeros. We will see that  $a_1$  can be written in terms of the zero-value time constants and low-frequency transfer constants evaluated with one reactive element infinite-valued at a time. The line of reasoning is the generalized version of the one used in determination of the zero frequency of a system with a single energy-storing element in Section 4.2. we will need to determine the  $\alpha_1^i$  coefficients in (4.33) first. This can be done by evaluating the low-frequency transfer constants when  $C_i$  is taken to infinity (short-circuited) while the other elements are still at zero value (opened capacitor and shorted inductor).

When  $C_i \rightarrow \infty$  in (4.34), (Figure 4.44) the transfer function from the input to output reduces to:

$$H^i \equiv H|_{\substack{C_i \rightarrow \infty \\ C_j=0 \\ i \neq j}} = \frac{\alpha_1^i}{\beta_1^i} \quad (4.44)$$

where  $H^i$  is the low-frequency transfer constant between the input and the

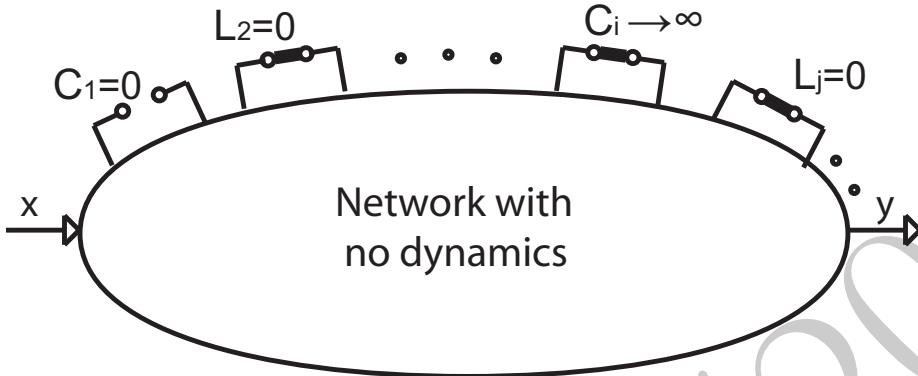


Figure 4.44: Calculation of  $H^i$  in an  $N$ -port with  $C_i$  shorted (infinite valued) and all other inductors and capacitors zero valued.

output with the reactive element  $i$  at its infinite value (i.e., short circuited capacitor or open circuited inductor)<sup>39</sup> and all others zero-valued. Since the value of  $H^i$  is frequency independent, it can be easily evaluated using the low frequency calculation methods developed earlier in Chapter 3. Noting that we have already determined  $\beta_1^i$  to be  $R_i^0$  in (4.36), we simply note that  $\alpha_1^i = R_i^0 H^i$  and hence according to (4.33) we have:

$$a_1 = \sum_{i=1}^N \alpha_1^i C_i = \sum_{i=1}^N R_i^0 C_i H^i = \sum_{i=1}^N \tau_i^0 H^i \quad (4.45)$$

which is stated for the case when all the energy-storing elements are capacitors. ▼ Result ▼

In the general case, where elements could be either capacitors or inductors, we can state  $a_1$  as

$$a_1 = \sum_{i=1}^N \tau_i^0 H^i \quad (4.46)$$

which is the sum of the products of zero-value time constants given by (4.35) or (4.38) and the first order transfer constants,  $H^i$ , evaluated with the energy storing element at the port  $i$  at its infinite value (i.e., shorted capacitors and open inductors), as shown in Figure 4.44. Note that the  $\tau_i^0$  time constants have already been computed in determination of  $b_1$  and hence all that needs to be calculated to determine  $a_1$  are  $H^i$  coefficients that can be evaluated using low-frequency (dc) calculations with the energy-storing element at port  $i$  infinite-valued (shorted capacitor or opened inductor).

<sup>39</sup>A comment on the notation is in order. As in Section 4.2, we place the index(es) of the infinite valued element(s) in the superscript. An index 0 in the superscript (as in  $R_1^0$ ) simply indicates that no reactive element is infinite valued, i.e., all elements are at their zero values.

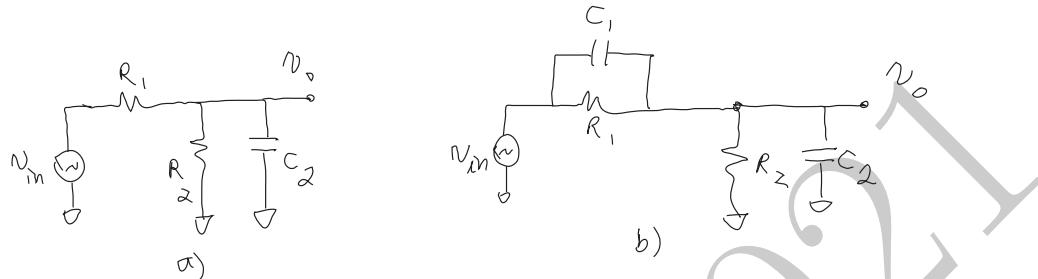


Figure 4.45: a) Low-pass model of amplifier input, b) pole-zero cancellation in an  $RC$  network.

Equation (4.46) suggests that if *all*  $H^i$  terms are zero, there will be *no* zeros in the transfer function. We will see in Section 4.6 that this is generally true. This suggests an easy test to determine whether there is a zero in the transfer function by looking for capacitors shorting of which (and inductors opening of which) results in a non-zero transfer constant<sup>40</sup>.

#### Example 4.4.1 Pole-Zero Cancelation: Oscilloscope Probe:

*It often occurs that a first stage drives a second one the presents a capacitive input impedance. in parallel with its input resistance, as shown symbolically in Figure 4.45a. It is often desirable for this stage not to impose any extra bandwidth limitation on the bandwidth. However, the circuit of Figure 4.45a has a low-pass response with a real pole at  $p = -1/(R_1 \parallel R_2)C_2$  with no zeros in the transfer function since shorting  $C_2$  results in a zero transfer function based on the above discussion. Interestingly, the same argument suggests that we may be able to introduce a zero in the transfer function if we create a situation where we have a non-zero transfer constant with shorting of a capacitor. One way to do so is to introduce a second capacitor  $C_1$  in parallel with  $R_1$  (assuming it is accessible), as shown in Figure 4.45b, introduces a LHP zero since the zero and infinite value transfer constants have the same polarity. Let us calculate  $b_1$  and  $a_1$ . To do so, let us calculate the  $\tau^0$ 's and  $H$ 's. We can easily see that:*

$$\tau_1^0 = (R_1 \parallel R_2)C_1$$

$$\tau_2^0 = (R_1 \parallel R_2)C_2$$

*The zero-value transfer constant is*

$$a_0 = H^0 = \frac{R_2}{R_1 + R_2}$$

---

<sup>40</sup>The above statement is valid in general. However, sometimes the transfer function has a pole that exactly coincides with a zero. When that happens the above procedure still predicts the existence of a zero. For example, this happen in the pathological case, where there is a resistor to ground in parallel with a capacitor that is not connected to the rest of the circuit! shorting of the capacitor does not change the low-frequency response of the circuit, hence predicting a zero. However, there is an uncoupled pole associated with the  $RC$  time-constant of the isolated circuit that coincides with the zero exactly.

We also have

$$\begin{aligned} H^1 &= 1 \\ H^2 &= 0 \end{aligned}$$

Now if we were trying to minimize the effect of  $b_1$  on the gain reduction at higher frequencies, we should force it to be equal to the equivalent coefficient in the numerator, namely,  $a_1/a_0$ . We can easily see that

$$\begin{aligned} b_1 &= (R_1 \parallel R_2)(C_1 + C_2) \\ \frac{a_1}{a_0} &= R_1 C_1 \end{aligned}$$

For  $b_1 = a_1/a_0$  to hold, we must have

$$b_1 = \frac{a_1}{a_0} = R_1 C_1 = R_2 C_2$$

In fact, due to the capacitive loop formed by  $C_1$  and  $C_2$  (when the input is nulled), there is only one pole and the circuit is a first order one. Hence the transfer function for arbitrary choice of component values is indeed completely determined by  $a_0$ ,  $a_1$ , and  $b_1$  to be

$$H(s) = a_0 \cdot \frac{1 + \frac{a_1}{a_0}s}{1 + b_1 s} = \frac{R_2}{R_1 + R_2} \cdot \frac{1 + R_1 C_1 s}{1 + (R_1 \parallel R_2)(C_1 + C_2)s}$$

which simply reduces to  $H^0 = R_2/(R_1 + R_2)$  if  $R_1 C_1 = R_2 C_2$  and behaves like a resistive divider.

It is noteworthy that had we stopped at  $b_1$  calculations using ZVT, we would have predicted a substantially different result for the response of the circuit in Figure 4.45b (a low pass one). Of course, that is because of the underlying assumption that there are no zeros in the approximation of the  $\omega_h$  with  $1/b_1$  in (4.42). In this case, as we can see we have a zero with a considerable effect that can be even made to completely cancel the pole.

Another way to see why this is happening is noting that when  $R_1 C_1 = R_2 C_2$  the impedances of the parallel combination formed by  $R_1$  and  $C_1$  has its pole at the same frequency as that of the impedance of the  $R_2$  and  $C_2$  parallel combination. Therefore, the voltage divider ratio reduces from  $Z_2/(Z_1 + Z_2)$  to  $R_2/(R_1 + R_2)$ .

The simple circuit of Figure 4.45b is used oscilloscope probes to minimize the effect of the input capacitance of the oscilloscope ( $C_2$ ), by introducing an intentional voltage divider by introducing  $R_1$  and adding a capacitor  $C_1$  in parallel. It turns out that recovering the lost gain is usually easier than correcting for the non-ideal transfer function, particularly in a time-domain measurement system that requires constant group delay (linear phase).

#### 4.4.1 ZVT Bandwidth Estimation for a System with Zeros

The bandwidth estimates of section 4.3.1 were based on several assumptions, one of which was absence of a zero in close proximity of  $\omega_h$ . However, this is not always the case as zeros may be present close to  $\omega_h$ . Example 4.3.7 is an instance of this where direct applications of (4.42) leads to a *gross* error in the bandwidth estimate. The approximation of (4.42) can be modified in the light of (4.46) to provide a more accurate estimate for  $\omega_h$ . Using a similar argument used to arrive at (4.42), we conclude that in the vicinity of  $\omega_h$ , the transfer function can be estimated as:

$$H(s) \approx a_0 \cdot \frac{1 + \frac{a_1}{a_0}s}{1 + b_1 s} \quad (4.47)$$

where according to (4.46)

$$\frac{a_1}{a_0} = \sum_{i=1}^N \tau_i^0 \frac{H^i}{H^0} \quad (4.48)$$

The transfer function of (4.47) is a first order system with a pole at  $-1/b_1$  and a zero at  $-a_0/a_1$ . The zero has the opposite effect on the magnitude of the transfer function compared to the pole since it *increases* the magnitude of the transfer function with frequency. First, let us assume that all  $H^i/H^0$  terms are positive. In this case, the numerator's first order coefficient,  $a_1/a_0$ , will be positive and the dominant zero is LHP. In this case, the  $\omega_h$  estimate can be modified to

$$\omega_h \approx \frac{1}{b_1 - \frac{a_1}{a_0}} = \frac{1}{\sum_{i=1}^N \tau_i^0 (1 - \frac{H^i}{H^0})} \quad (4.49)$$

which reduces to (4.42) when there are no zeros, i.e., all  $H^i$  terms are zero (corresponding to  $a_1 = 0$ ).

If some of the  $H^i/H^0$  terms are negative, it means that the transfer function has RHP zeros. However, the RHP zeros have exactly the same effect on the amplitude as LHP ones although their phase response is the opposite. Since  $\omega_h$  is a amplitude dependent quantity, the a LHP zero at a given frequency should produce the exact same  $\omega_h$  as a RHP one at the same frequency. Therefore, in the presence of RHP zeros, a better approximation for  $\omega_h$  is

$$\omega_h \approx \frac{1}{\sum_{i=1}^N \tau_i^0 (1 - |\frac{H^i}{H^0}|)} \quad (4.50)$$

which subsumes (4.49) for LHP zeros.

Equation (4.50) suggest that we can improve the accuracy of the results obtained from the direct application of the ZVT method for bandwidth estimation by modifying those ZVT's of the elements, for which the infinite-value transfer constants are non-zero (e.g., capacitors, shorting of which does not make the

output zero). The ZVT's bandwidth estimation can be applied as before, by replacing the original ZVT time constants with a new effective time constant,

$$\tau_i^{0'} = \tau_i^0 \cdot \left(1 - \left|\frac{H^i}{H^0}\right|\right) \quad (4.51)$$

in (4.42). Note that the effective ZVT's can be negative.

**Example 4.4.2 Common-Emitter Stage with Input Zero:** The pole-zero cancelation concept presented in the previous example can be applied to improve the frequency response of a common-emitter stage, as was shown earlier in Example 4.3.7. There we did not explain how we arrived at the value of  $C_1$ , but now considering the insight from the previous example, it is clear that  $C_1$  should be chosen so that the product,  $C_1 R_1$ , is equal to the product of the resistance seen looking into the base, i.e.,  $r_\pi$  and the equivalent capacitance in parallel with it. We know that the input equivalent capacitance is approximately  $C_t \equiv C_\pi + C_\mu(1 + g_m R_2)$  using the Miller approximation. For the values given,  $C_t = 1.72\text{pF}$ . Hence,

$$C_1 = C_t \cdot \frac{r_\pi}{R_1} = 4.3\text{pF}$$

which is the value used in Example 4.3.7. To calculate  $a_1$  we need the four time constants that have been computed in Examples 4.3.3 and 4.3.7 and the low-frequency transfer constants:

$$\begin{aligned} H^\pi &= 0 \\ H^\mu &= \frac{\alpha r_m \parallel R_2}{R_1 + \alpha r_m \parallel R_2} \approx \frac{r_m}{R_1} \approx 0.025 \\ H^L &= 0 \\ H^1 &= -g_m R_2 = -80 \end{aligned}$$

where  $H^\mu$  is the same as Example 4.2.5 leading to Equation (4.22) with the difference that in a BJT,  $r_m$  is replaced with  $\alpha r_m$  to account for the base current. Determination of  $H^1$  (which is the only  $H$  coefficient that really matters in this case) is straightforward, as it is simply the gain without the input voltage divider ( $C_1$  shorted). Since  $H^\mu$  and  $H^1$  are non-zero, the two ZVT's that need to be modified are (with  $H^0 = -57$ ):

$$\begin{aligned} \tau_\mu^{0'} &= \tau_\mu^0 \cdot \left(1 - \left|\frac{H^\mu}{H^0}\right|\right) = \tau_\mu^0 \cdot \left(1 - 0.0004\right) \approx \tau_\mu^0 \\ \tau_1^{0'} &= \tau_1^0 \cdot \left(1 - \left|\frac{H^1}{H^0}\right|\right) = \tau_1^0 \cdot \left(1 - 1.4\right) = -0.4 \cdot \tau_1^0 \end{aligned}$$

As can be seen, the modification to  $\tau_\mu^0$  is negligible, while the modified  $\tau_1^0$  has a significant impact.

The new bandwidth estimate using the effective time constants is:

$$\omega_h \approx \frac{1}{\sum_{i=1}^N \tau_i^{0'}} \approx \frac{1}{70\text{ps} + 1,200\text{ps} + 400\text{ps} - 1,230\text{ps}} \approx 2\pi \cdot 362\text{MHz}$$

♦ Numerical Example ♦

which is much closer to the SPICE results of  $f_h = 482MHz$ , shown in Figure 4.40, than the estimate of  $34MHz$  obtained in Example 4.3.7 using the unmodified ZVT's <sup>41</sup>.

As we can see after the correction, it is the time constants associated with  $C_\mu$  and  $C_L$  in conjunction with  $R_2$  that become significant and determine the bandwidth.

One thing to note is that we can quickly verify whether or not we need to use the approximation of (4.49) or (4.42) simply suffices, by determining if setting any of the energy-storing elements to its infinite value results in a non-zero transfer constant, namely if we have any non-zero  $H^i$  terms. For non-zero  $H^i$  we should see whether or not  $|H^i/H^0|$  can be ignored when compared to unity and see if its inclusion has a considerable effect on  $b_1$ . If that is the case, it should be subtracted from  $b_1$  and otherwise simply ignored.

#### 4.4.2 Creation of Zeros

Online YouTube lectures:

[Creation of zeros: series RC branch in parallel](#)

[Creation of zeros: parallel signal paths](#)

Unlike poles that are natural frequencies of the circuit and hence are not affected by the choice of the input and output variables<sup>42</sup>, zeros are a direct function of this choice, as evident by the presence of  $H^i$  terms in  $a_1$  (and all higher order numerator terms, as will be seen later.)

There will be zeros in the transfer function whenever at least one of the infinite-value transfer constants, namely,  $H^i$  is not zero. While this is a correct statement in general, we can gain more insight about the connection between the circuit configurations and the properties of the zeros by looking at a few scenarios that commonly occur.

<sup>41</sup>Another way to see this is by using (4.46) to find

$$\begin{aligned} a_1 &= \sum_{i=1}^N \tau_i^0 H^i = \tau_1^0 H^1 + \tau_\mu H^\mu = H^0(R_1 C_1 - r_m C_\mu) \\ &= -245,600ps + 40ps \approx -245.6ns \end{aligned}$$

Factoring  $a_0 = H^0 = -57$  out of the numerator, we have

$$\begin{aligned} H(s) &= a_0 \cdot \frac{1 + \frac{a_1}{a_0} s + \dots}{1 + b_1 s + \dots} = H^0 \cdot \frac{1 + (r_\pi C_t - r_m C_\mu)s + \dots}{1 + [r_\pi C_t + R_2(C_\mu + C_L)]s + \dots} \\ &\approx -57 \cdot \frac{1 + 4.30 \times 10^{-9}s + \dots}{1 + 4.74 \times 10^{-9}s + \dots} \end{aligned}$$

Now we have  $b_1 - a_1/a_0 = R_2 C_L + (R_2 + r_m) C_\mu \approx R_2(C_L + C_\mu) = 440ps$ . Using (4.50), we estimate  $\omega_h$ ,

$$\omega_h \approx \frac{1}{b_1 - \frac{a_1}{a_0}} \approx \frac{1}{R_2(C_L + C_\mu)} = \frac{1}{440ps} = 2\pi \cdot 362MHz$$

<sup>42</sup>With the exception of the somewhat pathological cases, where certain natural frequencies are not observable. As a trivial example consider the output being the voltage of resistor to ground isolated from the rest of the circuit.

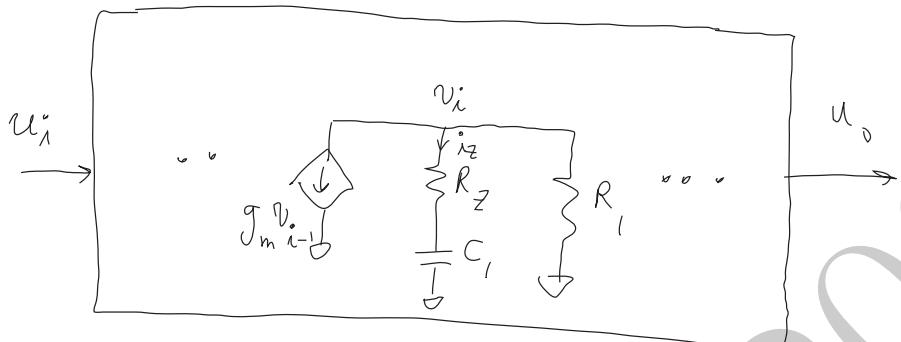


Figure 4.46: A series  $RC$  added from node  $i$  to ground in a multi-stage amplifier.

#### Example 4.4.3 A Series $RC$ to Ground:

Consider a generally multi-stage unilateral amplifier with a single feed-forward path whose  $i$ th stage is driven by a dependent current sources,  $g_m v_{i-1}$  and has a total resistance  $R_1$  to the ground, as shown in Figure 4.46. Let us show the low frequency gain of this amplifier as  $H^0$  and the transfer function as  $H_{\text{before}}(s)$ .

Now we introduce a series  $RC$  network with a resistance  $R_z$  and capacitance  $C_1$  and calculate the new transfer function,  $H_{\text{after}}(s)$  in terms of the old one,  $H_{\text{before}}(s)$ . Since the amplifier is unilateral, the dependent current source on node  $i$  disappears if the input is nulled and hence the only the resistance seen at low frequency is simply  $R_1$  to ground. Since no other reactive elements are directly connected to the node, shorting or opening of other energy storing elements has no impact on the resistance seen by  $C_1$  and thus the pole and the zero due to  $R_z$  and  $C_1$  will be uncoupled from the poles and zeros of  $H_{\text{before}}$ . For  $C_1$ , we have

$$\tau_z^0 = (R_z + R_1)C_1$$

$$H^z = H^0 \frac{R_z}{R_z + R_1}$$

and hence

$$H_{\text{after}}(s) = \frac{1 + R_z C_1 s}{1 + (R_z + R_1) C_1 s} \cdot H_{\text{before}}(s) \quad (4.52)$$

As we can see, the series  $RC$  network introduces a pole and a zero in the case of the unilateral single-path amplifier. In general, for an arbitrary network, addition of such an  $RC$  network introduces a pole-zero pair in the transfer function, but they won't be at the same frequencies predicted in the unilateral case due to the coupling among various energy-storing elements. The series network is an open at low frequencies hence the  $H^0$  does not change. Also it does not reduce to a short circuit and hence under the normal circumstances it does not diminish the transfer function for  $C_1 \rightarrow \infty$ , which implies  $H^z \neq 0$  that in turn means there will be a zero.

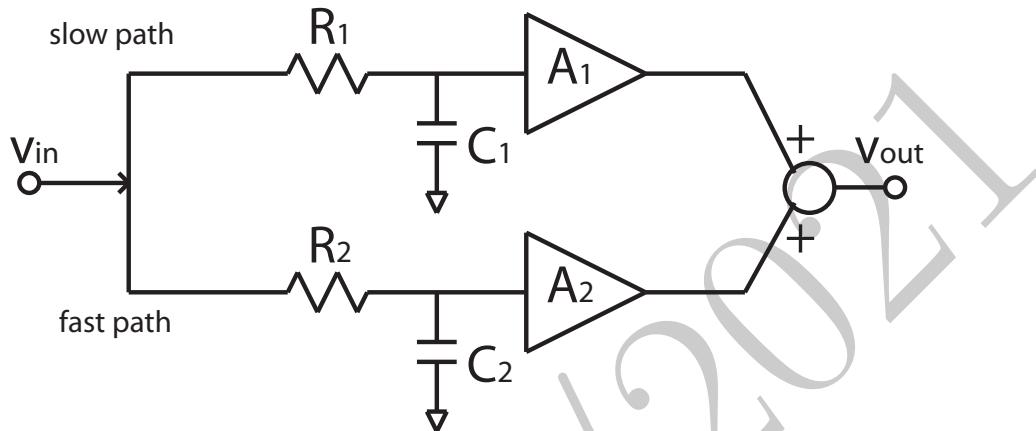


Figure 4.47: A system consisting of two signal paths each with a first order response followed by unilateral ideal voltage amplifiers.

*Another way to arrive at the same result is to find the frequency  $s$  at which  $v_i = 0$  irrespective of the branch current  $i_z$ . For this to happen we should have:*

$$v_x = R_z i_x + \frac{1}{C_1 s} i_x = 0$$

*which indicates,  $z = -1/R_z C_1$ .*

Yet another very common situation that gives rise to zeros is when there are two parallel forward paths with different dynamics. A simple example of this is given next

**Example 4.4.4 Dual-Path Zero Creation:** The system shown in Figure 4.47 consists of two first-order low-pass RC paths with single poles at  $p_1 = -1/R_1 C_1$  and  $p_2 = -1/R_2 C_2$ , respectively. The inputs are driven with an ideal voltage source. Neither path has a zero in its transfer function. The amplifiers are unilateral ideal voltage amplifiers (i.e.,  $R_{in} = \infty$  and  $R_{out} = 0$ ), with voltage gains  $A_1$  and  $A_2$ .

The low frequency gain, is readily seen to be:

$$a_0 = H^0 = A_1 + A_2$$

The time constants for each capacitor are

$$\begin{aligned}\tau_1 &= R_1 C_1 \\ \tau_2 &= R_2 C_2\end{aligned}$$

that result in the following two first order transfer constants for the two parallel systems:

$$H_1(s) = \frac{A_1}{1 + \tau_1 s}$$

$$H_2(s) = \frac{A_2}{1 + \tau_2 s}$$

The transfer function of the compound system is the sum of the two transfer functions, i.e.,

$$H(s) \equiv \frac{v_o}{v_i} = H_1(s) + H_2(s) = H^0 \cdot \frac{1 + \tau_z s}{(1 + \tau_1 s)(1 + \tau_2 s)}$$

where the zero time constant,  $\tau_z$ , is

$$\tau_z = -\frac{1}{z} = \frac{A_2 \tau_1 + A_1 \tau_2}{A_1 + A_2} \quad (4.53)$$

It is easy to see that the poles are those of the individual first order systems. But more interestingly, a new real zero is created that did not exist in either of the constituting first-order systems. This is essentially due to the fact that the summation of the two different responses in the two parallel systems will result in the two responses canceling each other at a (generally complex) frequency.

We can assume without loss of generality that  $\tau_1 > \tau_2$ , i.e.,  $p_1$  occurs at a lower frequency than  $p_2$  or the first path is slower than the second one in this case<sup>43</sup>. A careful look at (4.53) indicates that if the gains of the two paths have the same polarities (i.e.,  $A_1 A_2 > 0$ ), the zero falls between the two poles, as (4.53) could be looked at as the weighted average of  $\tau_1$  and  $\tau_2$ .

On the other hand when the two paths have opposite low frequency gain polarities ( $A_1 A_2 < 0$ ), the zero could be either LHP or RHP. To have a RHP zero we must have

$$-\frac{\tau_1}{\tau_2} < \frac{A_1}{A_2} < -1$$

otherwise the zero will still be LHP. We will see later in Section ?? how the location of the zero affects the time-domain response of the system using the model of Figure 4.47.

The above example clearly shows that two parallel signal paths can cause new zeros in the transfer function of the overall system that did not exist in either path alone. This is essentially due to the fact that the summation of the two different responses in the two parallel systems will result in the two responses canceling each other at the frequency,  $s = z$ .

We will see an example of a pair of imaginary zeros (or more generally complex conjugate zeros) in Example 4.5.7 of Section 4.5.

---

<sup>43</sup>If this was not the case, we could relabel  $\tau_1$  as  $\tau_2$  and vice versa

### 4.4.3 The Time-Domain Effect of Zeros

In this subsection, we investigate the impact of zeros on the time-domain response of the system.

Let us consider a *second* order system with a zero and two *real* poles. The system is arbitrary except for the fact that the poles are assumed to be real<sup>44</sup>. The transfer function of such a system can be written as

$$H(s) = A_0 \cdot \frac{1 - \frac{s}{z}}{(1 - \frac{s}{p_1})(1 - \frac{s}{p_2})} \quad (4.54)$$

where  $A_0$  the low frequency gain of the system. We can assume  $|p_2| > |p_1|$  without loss of generality. The zero could be a LHP or a RHP one. The above transfer function can be written as the sum of partial fractions:

$$H(s) = \frac{A_1}{1 - \frac{s}{p_1}} + \frac{A_2}{1 - \frac{s}{p_2}} = \frac{A_1}{1 + \tau_1 s} + \frac{A_2}{1 + \tau_2 s} \quad (4.55)$$

where  $\tau_1 = -1/p_1$  and  $\tau_2 = -1/p_2$  are the pole characteristic times and we have  $\tau_1 > \tau_2$ . We can solve for  $A_1$  and  $A_2$  to find

$$A_1 = A_0 \cdot \frac{p_2}{z} \cdot \frac{z - p_1}{p_2 - p_1} \quad (4.56a)$$

$$A_2 = -A_0 \cdot \frac{p_1}{z} \cdot \frac{z - p_2}{p_2 - p_1} \quad (4.56b)$$

This system can be completely modeled using the dual-path system of Example 4.4.4, shown in Figure 4.47, where we assumed that  $\tau_1 > \tau_2$ , i.e., the upper path with the low-frequency gain  $A_1$  is slower than the lower one with the gain  $A_2$ . It has two first-order parallel paths each with a single pole transfer function and no zeros. While the two-path system of Figure 4.47 may appear quite idealized, it can be used to represent any second order system with two LHP real poles and a zero. The each of these poles would be at exactly the same frequency as those of the real poles of the original second order system.

The main point of this decomposition is to use the equivalent system of Figure 4.47 to gain insight about the step response, by noting that the overall system time-domain response is the sum of the time domain responses of the two first order single-pole systems. This is useful since we know that the step response of a system with transfer function of  $H(s) = A/(1 + \tau s)$  is simply an exponential,  $A(1 - e^{-t/\tau})$ .

The step response of the complete system depends on the relative values of  $A_1$  and  $A_2$ . We saw in Example 4.4.4 that when the gains of both paths have the same polarity ( $A_1 A_2 > 0$ ), we have a LHP zero that falls between the two poles,  $p_1$  and  $p_2$  on the real ( $\sigma$ ) axis. In this case, the response is the sum of two exponentials going in the same direction with two different time constants as shown in Figure 4.48. So there would be no overshoot or undershoot in the step

---

<sup>44</sup>We will deal with the case of complex poles in the next section.

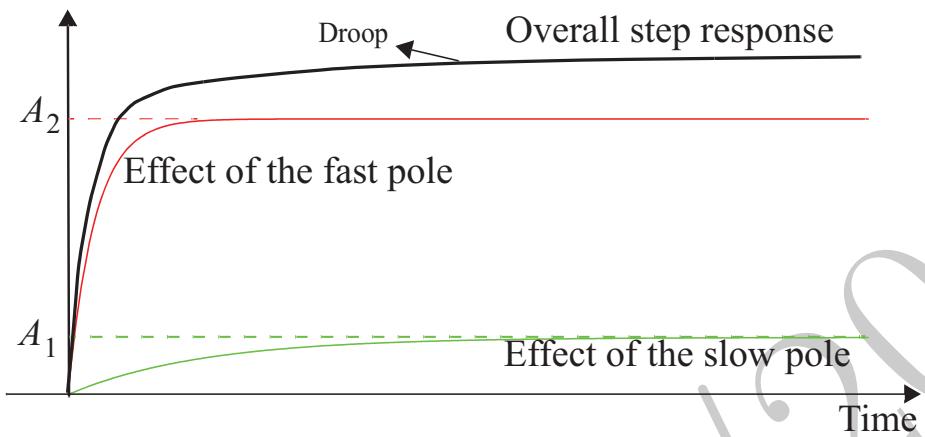


Figure 4.48: The step responses of two paths with same polarities and different time constants resulting in a *droop*.

response and the zero results in a *droop* in the step response, as the response associated with  $p_2$  settles quickly but the response corresponding to  $p_1$  takes a while longer to reach its final value.

We also saw in Example 4.4.4 that when the two gains have opposite polarities, we should look at the relative sizes of  $A_1$  and  $A_2$ . We saw that  $-1 < A_1/A_2 < 0$  results in a LHP zero closer to the origin than either  $p_1$  and  $p_2$ . Again we can determine the step response using the equivalent system of Figure 4.47. In this case, the two paths have opposite polarities and the magnitude of slower path's gain ( $A_1$ ) is smaller than the faster path ( $A_2$ ). Again without loss of generality assume that  $A_1$  is negative and  $A_2$  is positive<sup>45</sup>, as shown in Figure 4.49. We see that the faster path which has a higher gain results in an overshoot in the response that is eventually reduced by the slower path. It is noteworthy that the overshoot in this case is not the result of an under-damped response cause by a pair of complex conjugate poles with a high  $Q$  since the poles are real; rather it is caused by the zero in the transfer function.

Again from Example 4.4.4, we have a LHP zero if  $A_1/A_2 < -\tau_1/\tau_2$ , but the zero is at a higher frequency than either  $p_1$  and  $p_2$  and hence usually had a negligible effect. Another way to see this is by noting that in this case slow response has significantly higher gain than the faster one, so it modifies the slope of the response of the primary path slightly but its effect is completely diminished by the time the high-gain slower path reaches steady-state.

However, when the gains have opposite polarities and their ratio is in the range  $-\tau_1/\tau_2 < A_1/A_2 < -1$ , the faster response will have a more discernable effect on the response. In this case, we saw in Example 4.4.4 that we have a RHP zero. This time assume that  $A_1$  is positive and  $A_2$  is negative (again the

<sup>45</sup>The opposite assumption only inverts the sign of the step response.

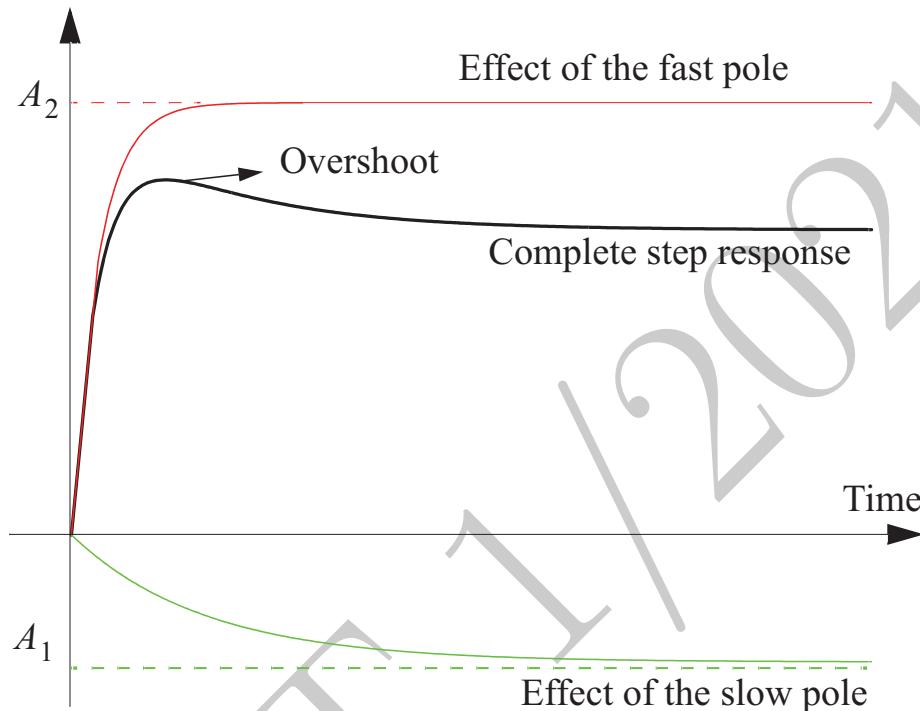


Figure 4.49: The step responses of two paths with opposite polarities and where  $|A_2| > |A_1|$  resulting in an overshoot.

opposite would only invert the response), as in Figure 4.50. The slower path still has a larger gain magnitude but the faster one has high enough gain to produce an *undershoot*. The undershoot is a trait associated with RHP zeros.

While we are on this topic there are a couple of special cases worth mentioning. One occurs when the zero is at the origin which happens when  $A_1 = -A_2$ . One typical example resulting in such a pole-zero constellation is an ac-couple amplifier which will have both low- and high-frequency cut-off frequencies. The overshoot (and/or undershoot) is taken to the extreme here. In this case,  $A_1$  and  $A_2$  will have the same magnitudes and opposite signs, hence the step response will overshoot first but eventually goes back to zero, as shown in Figure 4.51. This is consistent with the expected behavior of an ac-coupled amplifier, which cannot amplify the dc part of the step.

Another special case is when the zero coincides with the first pole,  $p_1$ . In this case, the zero cancels the effect of the first pole and the system essentially reduces to the fast path. This happens when  $A_1 = 0$ . This is can be thought of as the limiting cases of the droop (no droop) and overshoot (no overshoot).

To summarize, two parallel paths with the same polarity result in a real zero between  $p_1$  and  $p_2$  which causes a droop in the step response. Having two

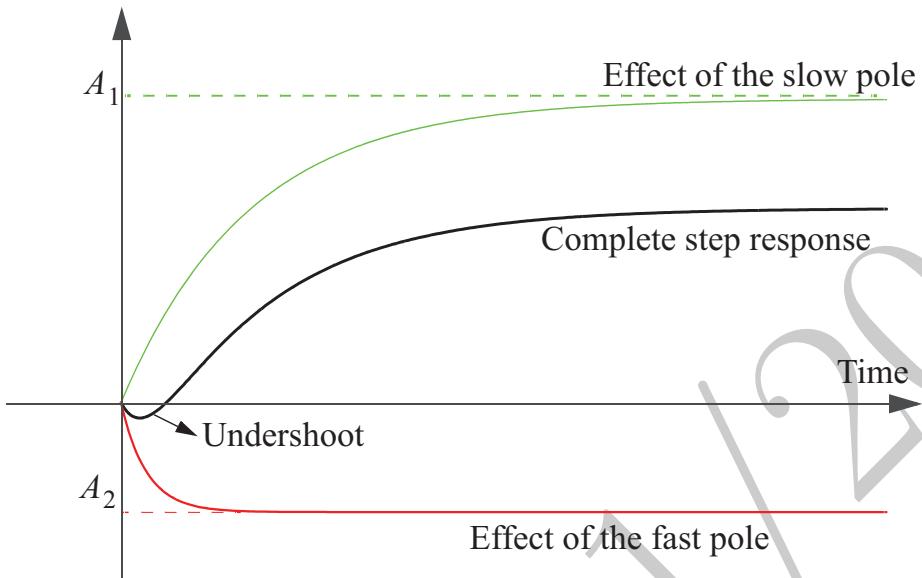


Figure 4.50: The step responses of two paths with opposite polarities and where  $|A_1| > |A_2|$  resulting in an undershoot.

signal paths with opposite polarities can result in an undershoot if the faster path has a smaller gain and an overshoot if it has a larger gain. These results are summarized in Figure 4.52.

The step response of a circuit is important in many applications. For example the time it takes for the output of an amplifier to a given accuracy can limit the useful speed of a high-resolution digital-to-analog converter. A droop is particularly bad, because the output gets close to its final value quickly, but does not quite get there to the necessary accuracy for a while.

## 4.5 Systems with Two Energy-Storing Elements

Online YouTube lectures:

[TTC Examples, 2nd Order System, Quality Factor and Natural Frequency](#)

As the next step in our development, it is instructive to focus on a system with two energy storage elements. The discussion in Section 4.6 is a superset of the material in this section. Nonetheless, this section provides an intermediate step toward TTC method and also elaborates on some of the properties of the second order systems, such as peaking, resonance, and the quality factor.

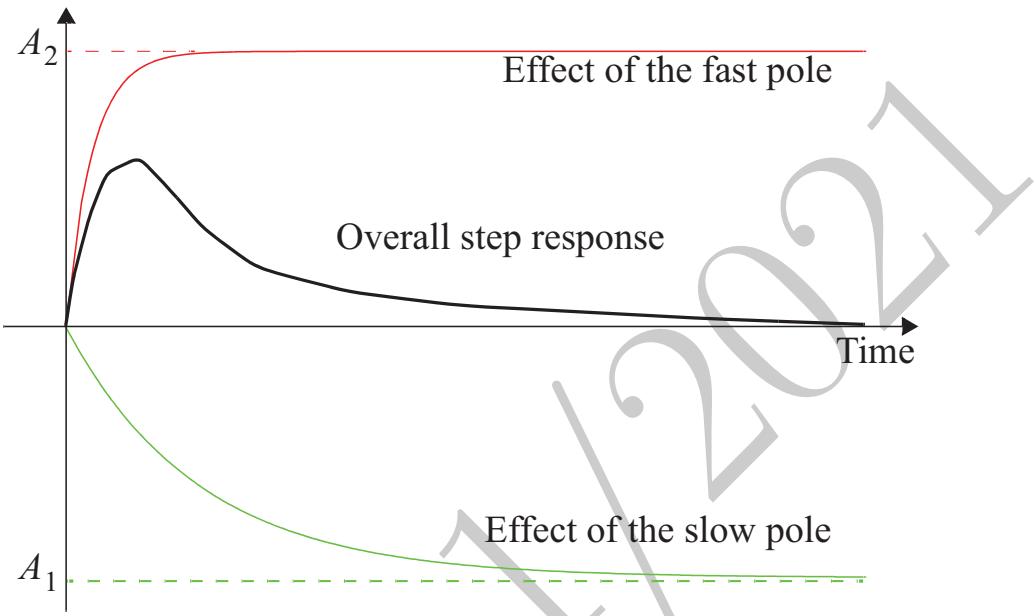


Figure 4.51: The step responses of two paths with opposite polarities and equal magnitudes,  $A_1 = -A_2$ , resulting in an extreme overshoot.

For a second order system, the transfer function of (4.1) reduces to

$$H(s) = \frac{a_0 + a_1 s + a_2 s^2}{1 + b_1 s + b_2 s^2} \quad (4.57)$$

Figure 4.53 shows a general representation of such a system where both these elements are capacitors, shown as  $C_1$  and  $C_2$ , attached to two different ports<sup>46</sup>. In general, each of the elements can be a capacitor or an inductor.

We have already determined  $a_0 = H^0$  to be the zero-value transfer constant. Coefficients  $b_1$  and  $a_1$  were determined for the more general case of  $N$ -energy-storage elements by (4.39) and (4.46). We will first focus on the case where both energy-storing elements are capacitors and then generalize to two arbitrary elements. Since coefficient  $s$  only appears a multiplicative term to  $C_1$  and  $C_2$ , the transfer function has the following form:

$$H(s) = \frac{H^0 + (H^1 R_1^0 C_1 + H^2 R_2^0 C_2) s + (\alpha_2^{12} C_1 C_2) s^2}{1 + (R_1^0 C_1 + R_2^0 C_2) s + (\beta_2^{12} C_1 C_2) s^2} \quad (4.58)$$

where coefficient  $\beta_2^{12}$  has units of ohms squared [ $\Omega^2$ ], and coefficient  $\alpha_2^{12}$  has the units of  $\Omega^2$  times the units of the transfer function. Since the numbering of  $C_1$

<sup>46</sup>If both elements happen to be in parallel (e.g., a parallel LC combination), we still use this representation assuming each element is connected to a port of its own.

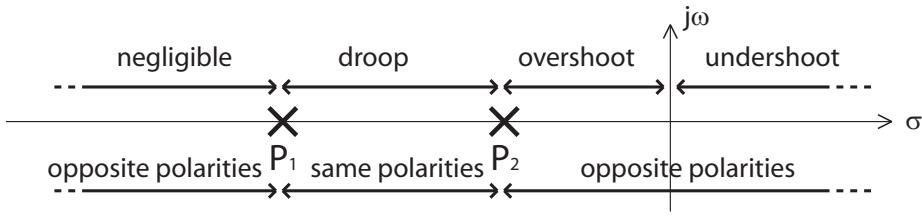


Figure 4.52: The impact of the location of the zero on the step response behavior and the implied relative polarity of the paths in its two-path equivalent.

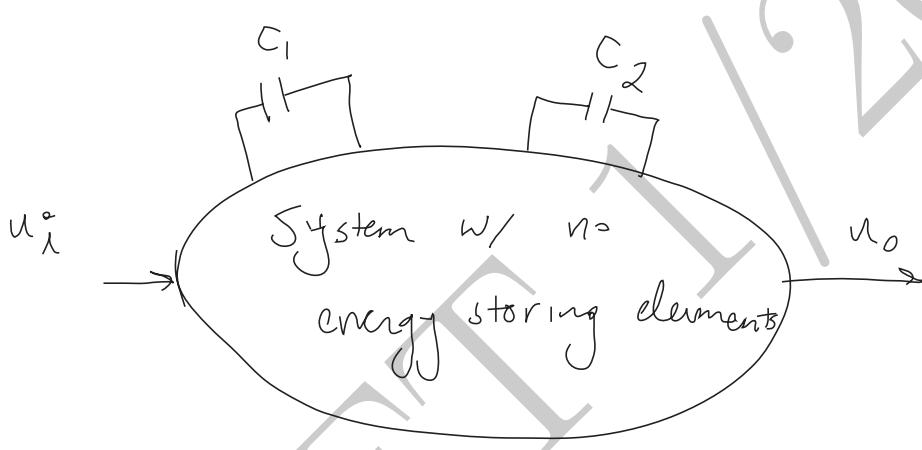


Figure 4.53: A system with two energy storing elements  $C_1$  and  $C_2$  at its ports.

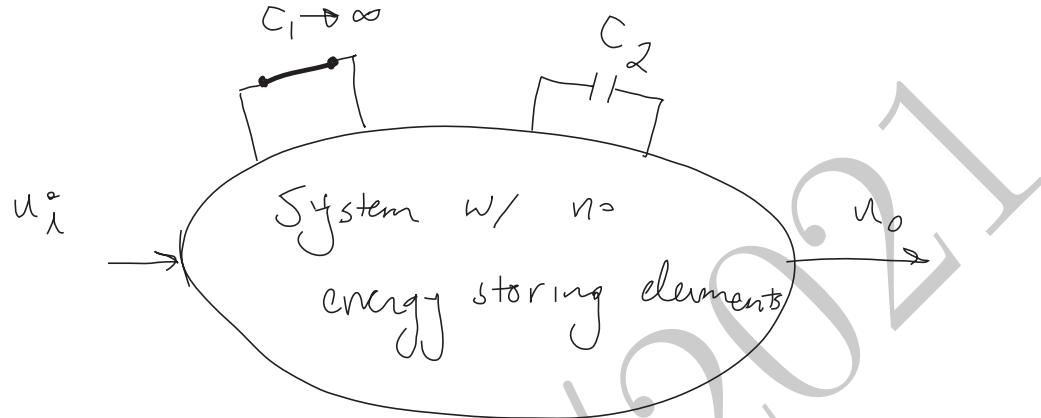
and  $C_2$  is arbitrary, (4.58) should be symmetrical with respect to the index. In other words, relabeling  $C_1$  as  $C_2$  and *vice versa* should not change the derived transfer function. Based on this we can conclude that  $\alpha_2^{12} = \alpha_2^{21}$  and  $\beta_2^{12} = \beta_2^{21}$ .

Now we try to determine  $\beta_2^{12}$  and hence  $b_2$  considering the case where  $C_1 \rightarrow \infty$ . In this case, the network reduces to one with a single energy-storing element, as shown in Figure 4.54. This network has a single time constant given by

$$\tau_2^1 = R_2^1 C_2 \quad (4.59)$$

where  $R_2^1$  is the resistance seen by  $C_2$  (the subscript) when  $C_1$  (the superscript) is infinite valued (shorted).

When  $C_1 \rightarrow \infty$ , the terms not containing  $C_1$  in the numerator and the denominator of (4.58) disappear as they are infinitesimal compared to the terms having a coefficient  $C_1$ . Therefore, (4.58) reduces to the transfer function of the

Figure 4.54: The system of 4.53 with  $C_1 \rightarrow \infty$  (short).

first-order system of Figure 4.54 given by:

$$H(s)|_{C_1 \rightarrow \infty} = \frac{C_1 s \cdot (H^1 R_1^0 + \alpha_2^{12} C_2 s)}{C_1 s \cdot (R_1^0 + \beta_2^{12} C_2 s)} = H^1 \cdot \frac{1 + \frac{\alpha_2^{12}}{H^1 R_1^0} C_2 s}{1 + \frac{\beta_2^{12}}{R_1^0} C_2 s} \quad (4.60)$$

Considering (4.59) and comparing (4.60) to (4.14), we can easily see that

$$\beta_2^{12} = R_1^0 R_2^1 \quad (4.61)$$

and hence

$$b_2 = R_1^0 C_1 R_2^1 C_2 = \tau_1^0 \tau_2^1 \quad (4.62)$$

As mentioned earlier, due to symmetry with respect to the choice of the indexes we have  $\beta_2^{12} = \beta_2^{21}$  and hence we conclude that

$$R_1^0 R_2^1 = R_2^0 R_1^2 \quad (4.63)$$

or equivalently

$$\tau_1^0 \tau_2^1 = \tau_2^0 \tau_1^2 \quad (4.64)$$

This is a useful result as in practice this provides a choice between two different ways to calculate the  $b_2$  coefficient, either by shorting  $C_1$  and determining the resistance seen by  $C_2$ , namely,  $R_2^1$ , or *vice versa* shorting  $C_2$  and calculating  $R_1^2$  that is the resistance seen by  $C_1$ . Often one is easier to calculate than the other.

Last, we determine  $a_2$  in (4.58) by setting both  $C_1$  and  $C_2$  to infinity (both short-circuited). In this case the second-order transfer constant will be

$$H^{12} \equiv H|_{\substack{C_1 \rightarrow \infty \\ C_2 \rightarrow \infty}} = \frac{\alpha_2^{12}}{\beta_2^{12}} \quad (4.65)$$

We have already calculated  $\beta_2^{ij}$  in (4.61), we find that  $\alpha_2^{12} = R_1^0 R_2^1 H^{12}$  and thus:

$$a_2 = R_1^0 C_1 R_2^1 C_2 H^{12} = \tau_1^0 \tau_2^1 H^{12} \quad (4.66)$$

where  $H^{12}$  is the low-frequency input-output transfer constant with the reactive elements 1 and 2 at their infinite value ( $C_1$  and  $C_2$  shorted).

Using (4.58) in conjunction with the calculated values for  $b_1$ ,  $a_1$ ,  $b_2$ , and  $a_2$  from (4.39), (4.46), (4.62), and (4.66), respectively, we determine the complete transfer function to be:

$$H(s) = \frac{H^0 + (\tau_1^0 H^1 + \tau_2^0 H^2)s + \tau_1^0 \tau_2^1 H^{12}s^2}{1 + (\tau_1^0 + \tau_2^0)s + \tau_1^0 \tau_2^1 s^2} \quad (4.67)$$

The above equation is in fact valid for both capacitors and inductors. The only difference is that for inductors, the time constants are simply given by

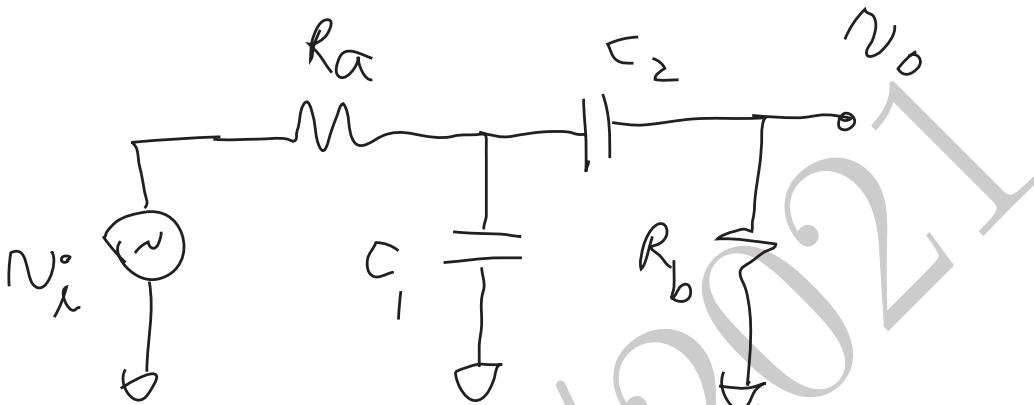
$$\tau_1^0 = \frac{L_1}{R_1^0} \quad (4.68a)$$

$$\tau_2^1 = \frac{L_2}{R_2^1} \quad (4.68b)$$

Now let us recapitulate. Equation (4.67) determines the transfer function of a system with two energy storage elements where,

1.  $H^0$  is the transfer constant of the system with all the elements at their zero values,
2.  $\tau_i^0 = R_i^0 C_i$  for a capacitor  $C_i$ , or  $\tau_i^0 = L_i / R_i^0$  for an inductor  $L_i$  is the zero-value time constant associated with the energy storage element  $i$ . It is determined by the resistance  $R_i^0$  which is the resistance seen by element  $i$  when the other element is zero valued (opened capacitor or shorted inductor),
3.  $\tau_j^i = R_j^i C_j$  for a capacitor  $C_j$  or  $\tau_j^i = L_j / R_j^i$  for an inductor  $L_j$  is a time constant associated with the energy storage element  $j$ . It is determined by the resistance  $R_j^i$  which is the resistance seen by element  $j$  when element  $i$  is infinite valued (shorted capacitor or opened inductor),
4. First-order transfer constants  $H^i$ 's are low-frequency transfer functions when element  $i$  is infinite valued (shorted capacitor or opened inductor) and the other one is zero valued, and
5. Second order transfer constant  $H^{ij}$  is the low frequency transfer function when both elements  $i$  and  $j$  are infinite valued (shorted capacitor or opened inductor).

The above approach will be generalized to the case of  $N$  energy-storage elements to determine all the  $a_i$  and  $b_j$  coefficients in (4.1) in Section 4.6.

Figure 4.55: A bandpass  $RC$  circuit.

**Example 4.5.1 Bandpass  $RC$  Filter** Consider the band-pass  $RC$  driven by an ideal voltage source, depicted in Figure 4.55. The low frequency gain is simply given by setting  $C_1$  and  $C_2$  to zero, leading to:

$$a_0 = H^0 = 0$$

We need to calculate  $R_1^0$ ,  $R_2^0$  and one of the  $R_1^2$  or  $R_2^1$  combinations. For completeness and as a demonstration of the validity of (4.63), we will calculate all four (we need three),

$$\begin{aligned} R_1^0 &= R_a \\ R_2^0 &= R_a + R_b \\ R_1^2 &= R_a \parallel R_b \\ R_2^1 &= R_b \end{aligned}$$

which are calculated by open- and short-circuiting of  $C_1$  and/or  $C_2$ . These coefficients allow us to calculate:

$$\begin{aligned} b_1 &= \tau_1^0 + \tau_2^0 = R_1^0 C_1 + R_2^0 C_2 = R_a C_1 + (R_a + R_b) C_2 \\ b_2 &= \tau_1^0 \tau_2^1 = R_1^0 R_2^1 C_1 C_2 = R_a R_b C_1 C_2 \end{aligned}$$

Note that we could calculate  $b_2$  as  $R_2^0 R_1^2 C_2 C_1$  which produces the same result.

Now for the  $H$  coefficients, we have:

$$H^1 = 0$$

$$H^2 = \frac{R_b}{R_a + R_b}$$

$$H^{12} = 0$$

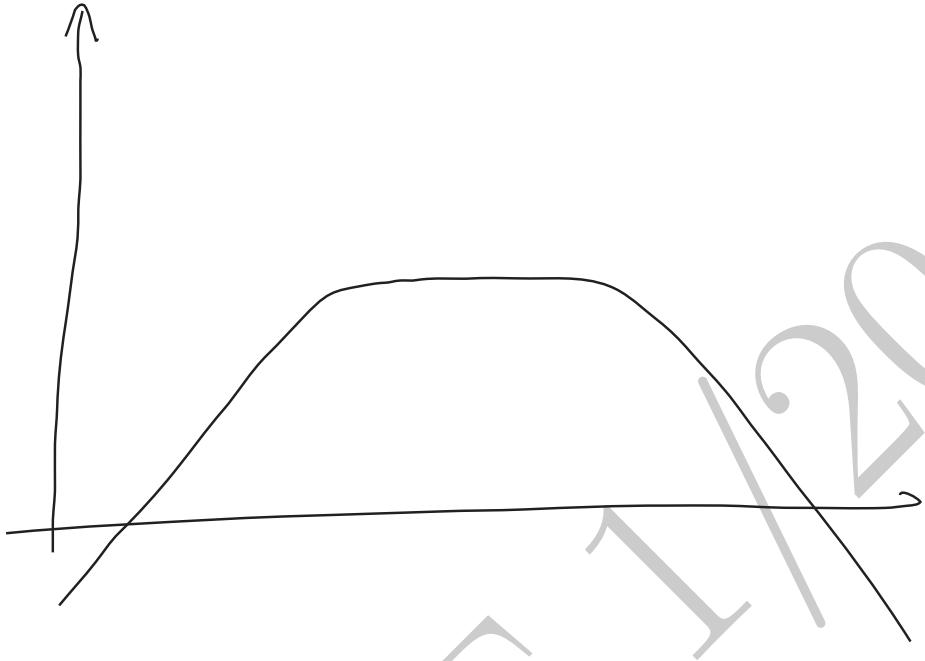


Figure 4.56: The transfer function of the bandpass  $RC$  filter of Figure 4.55.

which we can use to calculate,  $a_1$  and  $a_2$  to be:

$$\begin{aligned} a_1 &= \tau_1^0 H^1 + \tau_2^0 H^2 = R_1^0 C_1 H^1 + R_2^0 C_2 H^2 = R_b C_2 \\ a_2 &= \tau_1^0 \tau_2^1 H^{12} = R_1^0 R_2^1 C_1 C_2 H^{12} = 0 \end{aligned}$$

leading to the transfer function:

$$H(s) = \frac{R_b C_2 s}{1 + (R_a C_1 + R_a C_2 + R_b C_2)s + R_a R_b C_1 C_2 s^2}$$

The Bode plot of the transfer is shown in Figure 4.56.

**Example 4.5.2 Source Follower: Voltage Gain:** Consider the source follower of Figure 4.57. We looked at this in Example 4.2.3 with  $C_\pi$  alone. This time we look at it with both  $C_\pi$  and  $C_\mu$  present. The zero-value transfer constant is given by (4.16) to be:

$$a_0 = H^0 = \frac{R_2}{R_2 + r_m} = \frac{g_m R_2}{1 + g_m R_2}$$

The zero-value time constant associated with  $C_\pi$  was calculated in (4.17) to be

$$R_\pi^0 = \frac{R_1 + R_2}{1 + g_m R_2}$$

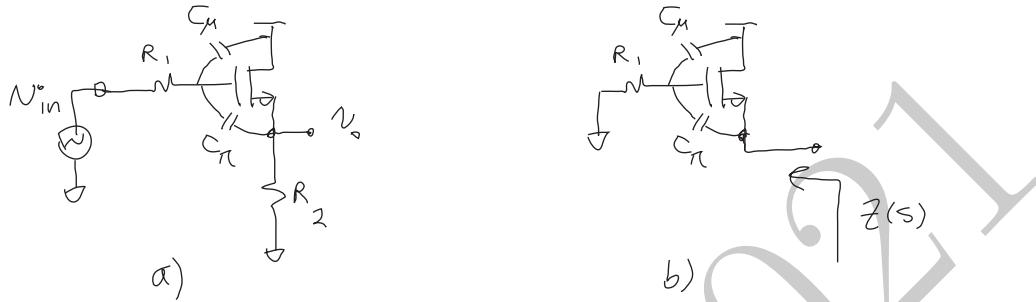


Figure 4.57: a) A source follower with both  $C_\pi$  and  $C_\mu$ , b) the model for output impedance calculations.

As for  $C_\mu$ , by inspection we have:

$$R_\mu^0 = R_1$$

Finally,  $R_\mu^\pi$  (the resistance seen by  $C_\mu$  when  $C_\pi$  is shorted) is

$$R_\mu^\pi = R_1 \parallel R_2$$

The transfer constants with  $C_\pi$ ,  $C_\mu$ , or both going to infinity (short-circuited) can be easily calculated by inspection to be:

$$H^\pi = \frac{R_2}{R_1 + R_2}$$

$$H^\mu = 0$$

$$H^{\pi\mu} = 0$$

Note that  $H^{\pi\mu}$  indicates that  $a_2 = 0$  due to (4.66). From the above we obtain:

$$\begin{aligned} b_1 &= R_1 C_\mu + \frac{R_1 + R_2}{1 + g_m R_2} C_\pi & a_1 &= \frac{R_2}{1 + g_m R_2} C_\pi \\ b_2 &= \frac{R_1 R_2 C_\pi C_\mu}{1 + g_m R_2} & a_2 &= 0 \end{aligned}$$

which allow use to express the transfer function exactly. When  $R_1 \ll R_2$  and  $r_m \ll R_2$  the transfer function reduces to:

$$H(s) \approx H^0 \cdot \frac{1 + r_m C_\pi s}{(1 + R_1 C_\mu s)(1 + r_m C_\pi s)} = \frac{R_2}{R_2 + r_m} \cdot \frac{1}{1 + R_1 C_\mu s}$$

which indicates that the zero and the pole introduced by the two parallel paths formed by the intrinsic source-follower and the capacitance  $C_\pi$  fall close to each other and cancel, making the transfer function a first order one controlled by the the  $R_1 C_\mu$  time constant. The zero associated with  $C_\pi$  is left-half plane since

the low frequency transfer function with  $C_\pi$  opened ( $H^0$ ) and shorted ( $H^\pi$ ) have the same signs. This is why it can diminish the pole. We see that ignoring this effect and relying on the  $b_1$  given by the ZVT approach alone can result in considerable underestimation of the bandwidth.

**Example 4.5.3 Source Follower: Output Impedance:** Let us revisit the output impedance of the source follower stage. We determined the output impedance and the output equivalent model for this stage with  $C_\pi$  as the only capacitor in Example 4.2.4. Now, we redo it with both  $C_\pi$  and  $C_\mu$ , as in Figure 4.57. Again, the output impedance is the parallel combination of the intrinsic output impedance  $Z(s)$  and the load resistance  $R_2$ . To determine  $Z(s)$  we calculate the time constant in a similar fashion as the previous example with  $R_2 \rightarrow \infty$  to be (still we only need three but completeness compute all four)

$$\begin{aligned}\tau_\pi^0 &= r_m C_\pi \\ \tau_\mu^0 &= R_1 C_\mu\end{aligned}$$

$$\begin{aligned}\tau_\pi^\mu &= r_m C_\pi = \tau_\pi^0 \\ \tau_\mu^\pi &= R_1 C_\mu = \tau_\mu^0\end{aligned}$$

Also, the transfer constants are:

$$\begin{aligned}Z^0 &= r_m = a_0 \\ Z^\pi &= R_1 \\ Z^\mu &= r_m \\ Z^{\pi\mu} &= 0\end{aligned}$$

Hence, the intrinsic output impedance can be written as

$$Z(s) = \frac{r_m + (\tau_\pi^0 Z^\pi + \tau_\mu^0 Z^\mu)s}{(1 + \tau_\pi^0 s)(1 + \tau_\mu^0 s)} = r_m \cdot \frac{1 + R_1(C_\pi + C_\mu)s}{(1 + R_1 C_\mu s)(1 + r_m C_\pi s)}$$

Comparing this to (4.19) we observe that the zero occurs at a slightly (but not substantially) lower frequency due to  $C_\mu$ . We also notice an additional pole at  $p_1 = -1/R_1 C_\mu$  which could at a considerably lower frequency than  $p_2 = -g_m/C_\pi \approx \omega_T$ , so the inductive range can be much smaller than predicted in Example 4.2.4 for  $R_1 \gg r_m$ , as we can see in the transfer function of Figure 4.58. The second pole eventually makes the output look capacitive beyond  $p_2 = -g_m/C_\pi$ .

This behavior is similar to that of the equivalent circuit, shown in Figure 4.59, which is similar to that of Example 4.2.4 with extra parallel capacitance. Assuming  $R_1 \gg r_m$ , the parameters of the equivalent model of Figure 4.59 can be calculated for the equivalent model to fit the common-source's output impedance.

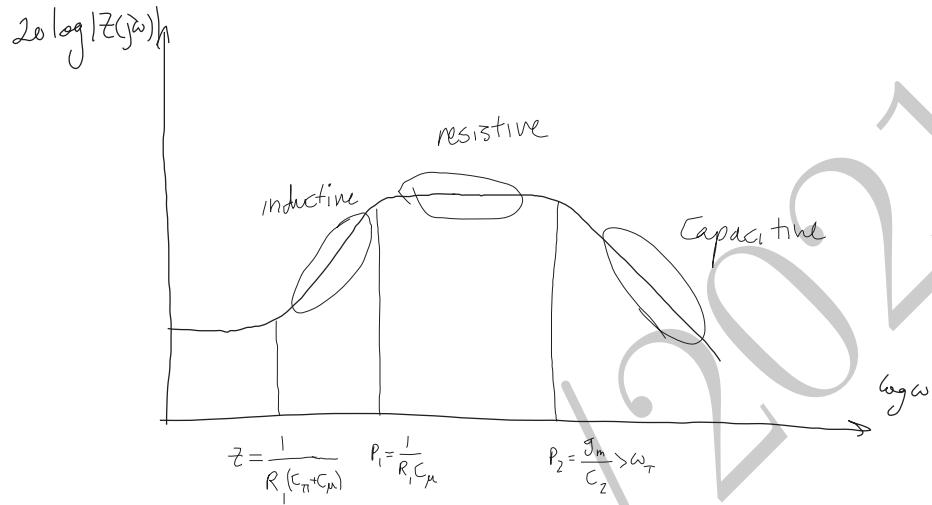


Figure 4.58: The output impedance of the source-follower of Figure 4.57b with both  $C_{\pi}$  and  $C_{\mu}$ .

These parameters are<sup>47</sup>:

$$L = r_m R_1 C_{\pi} \left(1 + \frac{C_{\pi}}{C_{\mu}}\right)^2$$

$$C = C_{\pi} \parallel C_{\mu}$$

$$R_a = R_1 \parallel r_m \left(1 + \frac{C_{\mu}}{C_{\pi}}\right)$$

$$R_b = r_m \left(1 + \frac{C_{\pi}}{C_{\mu}}\right)$$

for  $R_1 \gg r_m$  case<sup>48</sup>.

**Example 4.5.4 Input Impedance of Source Follower with Capacitive Load:** Consider the source follower of Figure 4.60a, driving a capacitive load,

<sup>47</sup>Note that we use the notation  $A \parallel B = AB/(A + B)$  as an operator. In the case of capacitors it corresponds to the *series* combination.

<sup>48</sup>It is possible to fit the model of Figure 4.59 to the output impedance of the stage *exactly*. In that case, the parameters of the equivalent model will be given by

$$L = \frac{(C_{\pi} + C_{\mu})^3 R_1^2 r_m}{C_{\pi}[C_{\mu} R_1 + C_{\pi}(R_1 - r_m)]} \quad R_a = \frac{(C_{\pi} + C_{\mu})^2 R_1 r_m}{C_{\pi} C_{\mu} R_1 + C_{\mu}^2 R_1 + C_{\pi}^2 r_m}$$

$$C = C_{\pi} \parallel C_{\mu} \quad R_b = \frac{(C_{\pi} + C_{\mu})^2 R_1 r_m}{C_{\pi}[C_{\mu} R_1 + C_{\pi}(R_1 - r_m)]}$$

These relations could be useful if one tries to modify the output impedance (e.g., by adding extra capacitors in parallel with  $C_{\mu}$  or  $C_{\mu}$ ).

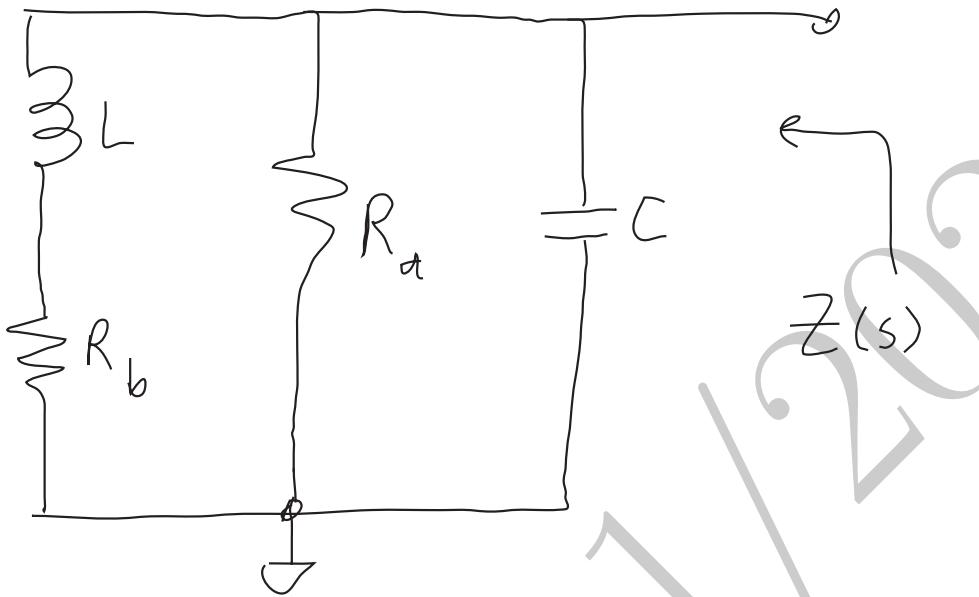


Figure 4.59: The output equivalent model of the source-follower of Figure 4.57b with both  $C_\pi$  and  $C_\mu$ .

$C_L$  (considering only  $C_\pi$  and  $C_L$ ). Calculating the input admittance,  $Y(s)$ , and inverting it is easier since the input impedance with both capacitors open is infinite. To calculate  $Y(s)$  we must drive the input with a voltage source (the stimulus) and take the input current as the output variable, as seen in Figure 4.60a.

First let us calculate  $Y^0$  when both capacitors are open. We simply have:

$$Y^0 = 0$$

In a similar way we have:

$$Y^\pi = 0$$

$$Y^L = 0$$

Now consider  $Y^{\pi L}$ . When both  $C_\pi$  and  $C_L$  are shorted, a short is seen looking into the input, and hence  $Y^{\pi L} = \infty$ . While correct this results in an indeterminate case, since for this configuration  $\tau_1^2 = \tau_2^1 = 0$  which results in a zero times infinity case for the  $a_2$  coefficient. This can be easily resolved by introducing a resistance,  $r_x$ , (which is always there anyway) in series with the  $C_L$  (or the input) and setting it to zero later, as shown in Figure 4.60b. The previously calculated  $Y^0$ ,  $Y^\pi$ , and  $Y^L$  terms are still zero. The new  $Y^{\pi L}$  is determined by inspection to be:

$$Y^{\pi L} = \frac{1}{r_x}$$

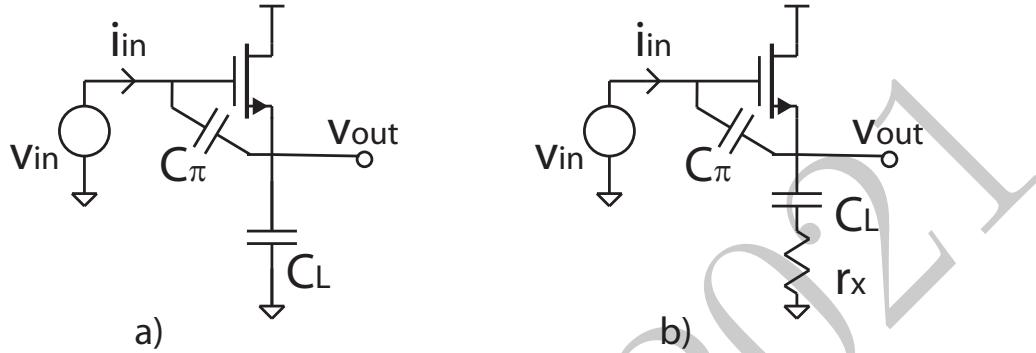


Figure 4.60: a) A source-follower with  $C_\pi$  driving a load capacitor  $C_L$ , b) the same stage with an infinitesimal resistance  $r_x$  in series with  $C_L$ .

Now to find the zero-value time constants, we see by inspection that

$$\begin{aligned}\tau_\pi^0 &= R_\pi^0 C_\pi = r_m C_\pi \\ \tau_L^0 &= R_L^0 C_0 = (r_m + r_x) C_L\end{aligned}$$

and finally we go ahead and calculate  $\tau_L^\pi$  as

$$\tau_L^\pi = R_L^\pi C_L = r_x C_L$$

Using these time and transfer constants and setting  $r_x \rightarrow 0$ , we obtain the input admittance:

$$Y(s) = \frac{r_m C_\pi C_L s^2}{1 + r_m (C_\pi + C_L) s}$$

Note that the above expression has a single pole, because of the arrangement of Figure 4.60 where  $C_\pi$  and  $C_L$  form a capacitive loop with the voltage source drive nulled (shorted). The above expression can be used to find the input impedance:

$$Z(s) = \frac{1}{Y(s)} = \frac{g_m}{C_\pi C_L s^2} + \frac{1}{(C_\pi \parallel C_L) s}$$

which has the model illustrated in Figure 4.61a and consists of the series combination of a capacitor and what is sometimes referred to as a “super capacitor” since it has a  $1/s^2$  behavior. More accurately it is a frequency dependent negative resistance (FDNR), as setting  $s = j\omega$  we see that it presents a negative resistance of

$$R = -\frac{g_m}{C_\pi C_L \omega^2}$$

at the input, as shown in Figure 4.61b. This can be useful in making oscillators or filters.

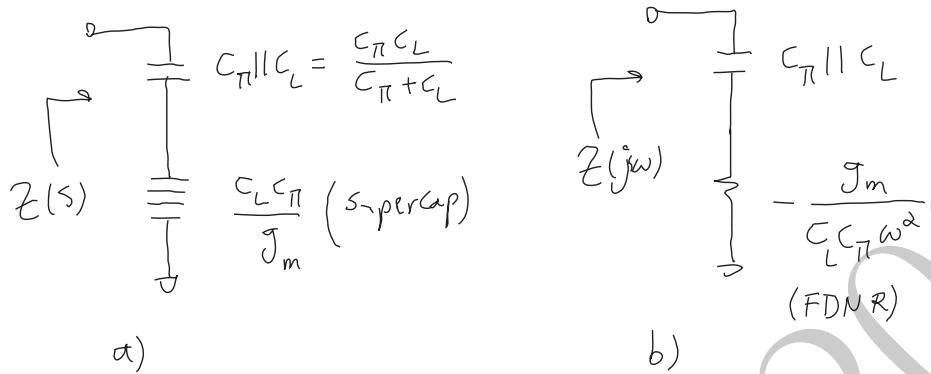


Figure 4.61: The input equivalent model of the source of Figure 4.60 as 1)an equivalent capacitor and a “super-capacitor” b) an equivalent capacitor and a frequency dependent negative resistance (FDNR).

Online YouTube lectures:

[TTC Examples, negative resistance, oscillator, RLC circuit](#)

**Example 4.5.5 Negative Resistance** In this example we analyze the cross-coupled NMOS pair connected across an RLC resonator, as shown in Figure 4.62, where biasing details are not shown<sup>49</sup>. Considering the input to be the differential current source,  $i_{in}$ , and the output to be the differential voltage,  $v_{out}$ , we determine the transfer function. The time constants are:

$$\begin{aligned}\tau_C^0 &= 0 & \tau_L^0 &= L(-g_m/2 + G_o) \\ \tau_C^L &= C/(-g_m/2 + G_o)\end{aligned}$$

All transfer constants with the exception of  $H^L$  are zero. Defining  $G_{eff} \equiv g_m/2 - G_o$ , we easily see that  $H^L = -1/G_{eff}$ . These time and transfer constants correspond to  $a_0 = 0$ ,  $a_1 = L$ , and  $a_2 = 0$ , as well as  $b_1 = -G_{eff}L$  and  $b_2 = LC$ . Hence, we can write the transfer function as

$$H(s) \equiv \frac{v_{out}}{i_{in}} = \frac{Ls}{1 - G_{eff}Ls + LCs^2} \quad (4.69)$$

As can be easily seen, for  $g_m/2 > G_o$  the denominator has a pair of RHP complex conjugate poles, corresponding to an exponentially growing response consistent with the start-up of a cross-coupled LC oscillator. This example shows that the TTC approach is applicable to both stable and unstable circuits.

<sup>49</sup>For instance the transistors could be biased through the center tap of the inductor

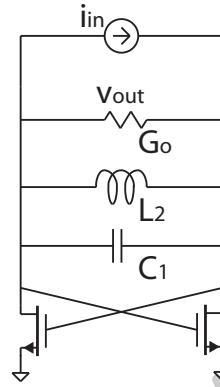


Figure 4.62: A negative resistance cross-coupled oscillator.

#### 4.5.1 Damping Ratio, Quality Factor, and Complex Poles

It is more common (and often more useful) to write the denominator of the transfer function of (4.57) in terms of the natural frequency,  $\omega_n$ , and the quality factor,  $Q$ ,

$$H(s) = \frac{N(s)}{1 + \frac{s}{Q\omega_n} + \frac{s^2}{\omega_n^2}} = \frac{N(s)}{1 + 2\zeta \frac{s}{\omega_n} + \frac{s^2}{\omega_n^2}} \quad (4.70)$$

where  $N(s)$  is the numerator in general, and  $Q = 1/2\zeta$  is a measure of the energy loss per cycle in the system ( $\zeta$  is called the damping ratio). Namely,

$$Q \equiv 2\pi \cdot \frac{\text{Energy Stored}}{\text{Energy Dissipated per Cycle}} = \omega \cdot \frac{\text{Energy Stored}}{\text{Power Dissipated}} \quad (4.71)$$

A simple comparison of (4.70) and (4.57) provides the following useful result:

$$\boxed{Q = \frac{1}{2\zeta} = \frac{\sqrt{b_2}}{b_1}} \quad (4.72)$$

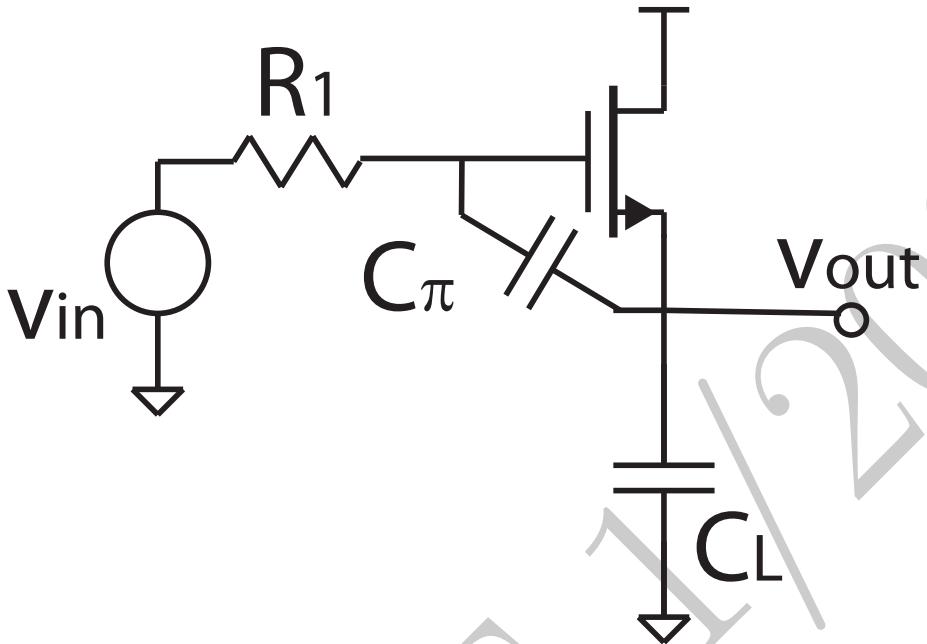
which can be written in terms on the time-constants for a second-order system as:

$$\frac{1}{Q} = 2\zeta = \frac{\tau_1^0 + \tau_2^0}{\sqrt{\tau_1^0 \tau_2^0}} = \sqrt{\frac{\tau_1^0}{\tau_2^0}} + \sqrt{\frac{\tau_2^0}{\tau_1^0}} \quad (4.73)$$

where (4.64) has been used in the last step to arrive at a more symmetrical result. It is easy to see from quadratic roots of the denominator of (4.70) that for  $Q > \frac{1}{2}$  the roots of the denominator become complex.

Also comparing (4.70) and (4.57), the undamped resonance or natural frequency,  $\omega_n$ , can be readily related to the  $b_2$  by

$$\omega_n = \frac{1}{\sqrt{b_2}} \quad (4.74)$$

Figure 4.63: A source follower stage driving a capacitive load,  $C_L$ .

which can be written in terms of the time constants as

$$\omega_n = \frac{1}{\sqrt{\tau_1^0 \tau_2^1}} = \frac{1}{\sqrt{\tau_2^0 \tau_1^2}} \quad (4.75)$$

Equations (4.72) and (4.74) are useful rules of thumb to remember particularly in the light of the relatively straightforward relation between  $Q$  and  $\omega_n$  with  $b_1$  and  $b_2$  coefficients given by (4.72). We will see later that it is useful in most practical systems of higher order to estimate the amplitude and the frequency of peaking of the response.

The maximum peaking frequency is given by

$$\omega_{peak} = \omega_n \sqrt{1 - 2\zeta^2} = \omega_n \sqrt{1 - \frac{1}{2Q^2}} \quad (4.76)$$

Next we see an example with substantial peaking. As mentioned earlier, the ZVT method is completely oblivious to such behavior as it provide no information about the imaginary part of the poles.

**Example 4.5.6 (Source Follower with Capacitive Load; Gain:)** Now we consider the gain of the source-follower stage with a source resistance  $R_1$  driving a capacitive load, as shown in Figure 4.63<sup>50</sup>. For now let us ignore  $C_\mu$  and

only consider  $C_\pi$  and  $C_L$ . The ZVTs are:

$$\tau_\pi^0 = r_m C_\pi \quad \tau_L^0 = r_m C_L$$

and  $\tau_\pi^L$  is given by

$$\tau_\pi^L = R_1 C_\pi$$

### ♦ Numerical Example ♦

Let assume  $C_\pi = C_L = 50fF$  and  $g_m = 20mS$  similar to example 4.3.6. This time assume an  $R_1 = 2k\Omega$ , we have,

$$\begin{aligned} b_1 &= \tau_\pi^0 + \tau_L^0 = 2.5ps + 2.5ps = 5ps \\ b_2 &= \tau_L^0 \tau_\pi^L = 2.5ps \times 100ps = 250(ps)^2 \end{aligned}$$

which indicates

$$\begin{aligned} Q &= \frac{\sqrt{b_2}}{b_1} = \sqrt{10} = 3.16 \\ \omega_n &= \frac{1}{\sqrt{b_2}} = 2\pi \cdot 10GHz \end{aligned}$$

As mentioned before in Chapter 2, the peaking occurs roughly around  $\omega_{peak}$  which is close to  $\omega_n$  for reasonably large values of  $Q$ . The peaking itself is approximately equal to  $Q$ , hence we should expect a peaking of approximately  $20\log_{10}(Q) = 10dB$  in the amplitude around 10GHz in the response. A SPICE simulations of the source follower with no  $C_\mu$  is shown in Figure 4.64, which clearly shows 10.2dB of peaking at 9.8GHz. The fact that  $Q$  is greater than 0.5 clearly indicates that we have a pair of complex conjugate poles. Also the simulated  $f_h$  is approximately 15.5GHz which is relatively close to the modified estimate of 15.9GHz obtained from (4.49) using the  $a_1/a_0$  term<sup>51</sup>. The original ZVT estimate of (4.42) predicts an  $f_h$  of twice as large at 32GHz.

As a final note, this peaking is usually attenuated by  $C_\mu$  and is not as pronounced as shown in this example. Nonetheless, the poles usually remain complex as the  $Q$  is often greater than 0.5.

**Example 4.5.7 Parallel LC in Series** In the common-source amplifier of Figure 4.65 shown together with its small-signal model we have introduced a parallel LC in series. If we ignore the transistor parasitic capacitors, the ZVTs

<sup>50</sup>This could be the case if the stage is biased with a current source.

<sup>51</sup>To calculate the  $a_i$  coefficients, we need the  $H^i$  terms that are:

$$\begin{aligned} H^0 &= 1 & H^L &= 0 \\ H^\pi &= 1 & H^{\pi L} &= 0 \end{aligned}$$

which predict

$$a_0 = 1 \quad a_1 = r_m C_\pi \quad a_2 = 0$$

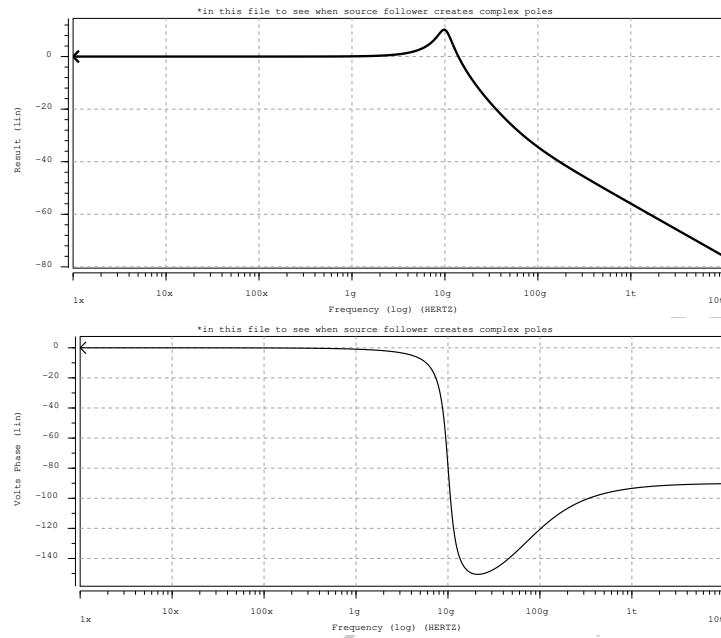


Figure 4.64: The amplitude and phase response of the voltage transfer function of the source-follower with a source resistance  $R_1 = 2k\Omega$ , driving a capacitive load, shown in Figure 4.63.

are

$$\begin{aligned}\tau_L^0 &= \frac{L}{R_1 + R_2} \\ \tau_C^0 &= 0\end{aligned}$$

Since  $\tau_C^0 = 0$  and  $\tau_C^0 \tau_L^C = \tau_L^0 \tau_C^L$  according to (4.64), we can avoid an indeterminant case<sup>52</sup> by calculating  $\tau_C^L = (R_1 + R_2)C$  and hence

$$\tau_L^0 \tau_C^L = LC$$

---

<sup>52</sup>Since  $\tau_C^0 = 0$  and  $\tau_L^C = L/0$ , the product is indeterminant. If one insists on using the product  $\tau_C^0 \tau_L^C$ , it can be determined by placing a resistor  $r_x$  in series with  $L$  and setting it to zero in the final result.

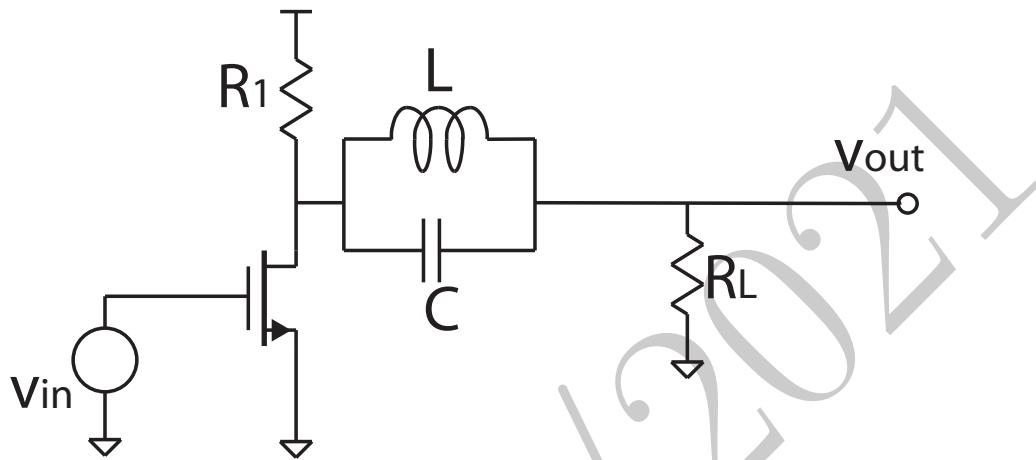


Figure 4.65: A common-source amplifier with a parallel LC trap in series, b) its small-signal model.

To calculate the numerator we need to determine the  $H$  coefficients, which are

$$\begin{aligned} H^0 &= -g_m(R_1 \parallel R_2) \\ H^L &= 0 \\ H^C &= -g_m(R_1 \parallel R_2) = H^0 \\ H^{LC} &= -g_m(R_1 \parallel R_2) = H^0 \end{aligned}$$

which result in

$$H(s) = H^0 \frac{1 + LCs^2}{1 + \frac{L}{R_1 + R_2}s + LCs^2}$$

where according to (4.72), we have

$$Q = (R_1 + R_2) \sqrt{\frac{C}{L}}$$

As can be seen from the transfer function, there is a pair of imaginary zeros at  $\pm j/\sqrt{LC}$ .

## 4.6 The Generalized Method of Time and Transfer Constants

Online YouTube lectures:

[Generalized Time- and Transfer constants; higher order terms](#)

## ▼ Derivation ▼

Any network with  $N$  reactive elements attached to it can be represented as an  $N$ -port with no frequency-dependent elements (e.g., containing only resistors and dependent voltage and current sources) and the reactive elements (namely inductors and capacitors) attached to its ports<sup>53</sup>, as shown in Figure 4.25.

The impedance of the capacitors,  $C_i$  and inductors  $L_j$ , are  $1/C_i s$  and  $L_j s$ , respectively. As can be seen, the only way the  $s$  coefficients can come about in the final transfer function is in conjunction with an  $L$  or a  $C$  as a multiplicative factor. In other words, in all subsequent network calculations, the original  $C_i$  or  $L_j$  associated with the  $s$  coefficient will accompany it through all calculations. Let us initially limit our discussion to capacitors and deal with the combined capacitor and inductor case later. Based on the above argument, the transfer function in its most general case must have the following form<sup>54</sup>:

$$H(s) = \frac{a_0 + (\sum_{i=1}^N \alpha_1^i C_i)s + (\sum_i \sum_{j < i}^{1 \leq j \leq N} \alpha_2^{ij} C_i C_j)s^2 + \dots}{1 + (\sum_{i=1}^N \beta_1^i C_i)s + (\sum_i \sum_{j < i}^{1 \leq j \leq N} \beta_2^{ij} C_i C_j)s^2 + \dots} \quad (4.77)$$

where  $\beta_k^{(\dots)}$  coefficients have units of  $\Omega^k$  and  $\alpha_k^{(\dots)}$  coefficients have units of  $\Omega^k$  times the units of the transfer function itself. Note that the sums are defined in such a way that for any two indexes  $m$  and  $n$  only one of the  $\beta_2^{mn}$  and  $\beta_2^{nm}$  is present in the sum to avoid multiple permutations of the same product<sup>55</sup>. Now, since a relabeling the capacitors should not change the poles and zeros of the transfer function and (4.77) should be valid for all values of  $C_i$  (including zero and infinity), we conclude that  $\beta_2^{mn} = \beta_2^{nm}$ . A similar argument can be applied to  $a_2$  coefficients in the numerator to conclude  $\alpha_2^{mn} = \alpha_2^{nm}$ . Also note that the higher order terms in (4.77) denoted by (...) have coefficients that are sums of products of at least three different capacitors.

The arguments of Sections 4.3 and 4.4 were made for a similar situation with  $N$  reactive elements hence (4.39) and (4.46) are still valid for determination of coefficients  $b_1$  and  $a_1$  in (4.1), respectively. This simply implies that  $\beta_1^i = R_i^0$  and  $\alpha_1^i = R_i^0 H^i$  in (4.77).

Now we determine higher order coefficients in (4.1). Next assume that we set  $C_i$  to infinity and consider a capacitor  $C_j$  at port  $j$  while all other capacitors have a value of zero (i.e., are open). The network will look like Figure 4.66. This is essentially another first-order system yet *different* from the one in Figure 4.26

<sup>53</sup>If more than one reactive element is connected to the same pair of terminals, each one of them is assumed to have a port of its own with a separate index.

<sup>54</sup>As mentioned in the derivation of the ZVT in section 4.3, the reason the higher order terms in (4.77) do not contain square of a single capacitor (e.g.,  $(C_i)^2$ ) is that such a term would produce a second order transfer function in (4.34) for the first order system of Figure 4.26 has only one energy storing element.

<sup>55</sup>More generally, we can expect that any circular rotation of the  $ijk\dots$  indexes in the superscript of  $\alpha_l^{ijk\dots}$  and  $\beta_l^{ijk\dots}$  should result in the same value due to the same invariance to the labeling of the capacitors.

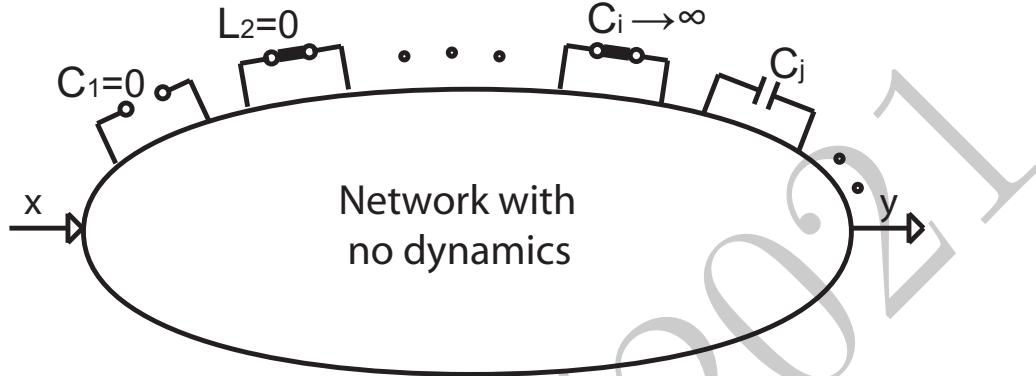


Figure 4.66: An  $N$ -port with a single capacitor  $C_j$  present, while  $C_i$  is infinite valued (shorted) and all the inductors and capacitors zero valued.

used to determine  $b_1$ . The time constant of this new first-order system is

$$\tau_j^i = R_j^i C_j \quad (4.78)$$

where  $R_j^i$  is the resistance seen at port  $j$  with port  $i$  shorted. Evaluating (4.77) with  $C_i \rightarrow \infty$  and all other capacitors other than  $C_i$  and  $C_j$  at their zero value (i.e., open) we obtain:

$$H(s)|_{C_i \rightarrow \infty} = \frac{C_i s \cdot (\alpha_1^i + \alpha_2^{ij} C_j s)}{C_i s \cdot (\beta_1^i + \beta_2^{ij} C_j s)} = \frac{\alpha_1^i}{\beta_1^i} \cdot \frac{1 + \frac{\alpha_2^{ij}}{\alpha_1^i} C_j s}{1 + \frac{\beta_2^{ij}}{\beta_1^i} C_j s} \quad (4.79)$$

which is the transfer function of the new first order system shown in Figure 4.66. Equating the coefficient of  $s$  in the denominator of (4.79) to (4.78), we obtain:

$$\beta_2^{ij} = \beta_1^i R_j^i = R_i^0 R_j^i \quad (4.80)$$

where we have used (4.36) in the last step. The second coefficient of the denominator,  $b_2$ , can be calculated as:

$$b_2 = \sum_i^{\text{1} \leq i < j \leq N} \sum_j R_i^0 C_i R_j^i C_j \quad (4.81)$$

▼ Intermediate Result ▼ which in the general case can be written as

$$b_2 = \sum_i^{\text{1} \leq i < j \leq N} \sum_j \tau_i^0 \tau_j^i \quad (4.82)$$

One important point is that since  $\beta_2^{ij} = \beta_2^{ji}$ , as discussed earlier, we obtain the following equality:

$$R_i^0 R_j^i = R_j^0 R_i^j \quad (4.83)$$

This equality can be concluded by a symmetry argument, noting that the same coefficient in (4.77) must be obtained by first computing the zero-value time constant of  $C_j$  and then shorting port  $j$  and determining the time constant associated with  $C_i$ . Equation (4.83) is important in practice since it provides alternative ways of calculating higher order time constant products, some of which may be more straightforward in the actual circuit, as will be seen in the subsequent examples.

Now to obtain  $a_2$ , we will let both  $C_i$  and  $C_j$  to go to infinity (short circuited) and all other reactive elements to be at their zero value (e.g., open capacitors). The the second-order input-output transfer constants are simply given by:

$$H^{ij} \equiv H|_{\substack{C_i, C_j \rightarrow \infty \\ C_k=0 \\ i \neq j \neq k}} = \frac{\alpha_2^{ij}}{\beta_2^{ij}} \quad (4.84)$$

Since we have already determined  $\beta_2^{ij}$  in (4.80), we determine that  $\alpha_2^{ij} = R_i^0 R_j^i H^{ij}$  and thus:

$$a_2 = \sum_i \sum_j^{1 \leq i < j \leq N} R_i^0 C_i R_j^i C_j H^{ij} \quad (4.85)$$

which again more generally can be written as

$$a_2 = \boxed{\sum_i \sum_j^{1 \leq i < j \leq N} \tau_i^0 \tau_j^i H^{ij}} \quad (4.86)$$

### ▼ Intermediate Result ▼

where the transfer constant,  $H^{ij}$ , is the low-frequency input-output transfer function with both ports  $i$  and  $j$  shorted (or in general the reactive elements at ports  $i$  and  $j$  at their infinite value). The above approach can be continued to determine higher order  $a_i$  and  $b_i$  coefficients applying induction to (4.77)

### ▼ Result ▼

In general,  $n$ th order  $b_n$  coefficient of the denominator is given by:

$$\boxed{b_n = \sum_i \sum_j^{1 \leq i < j < k \dots \leq N} \dots \tau_i^0 \tau_j^i \tau_k^j \dots} \quad (4.87)$$

and the  $a_n$  coefficient for the numerator is

$$\boxed{a_n = \sum_i \sum_j^{1 \leq i < j < k \dots \leq N} \dots \tau_i^0 \tau_j^i \tau_k^j \dots H^{ijk} \dots} \quad (4.88)$$

where  $\tau_k^{ijk\dots}$  corresponds to the time constant due to the reactive element at port  $k$  and the low frequency resistance seen at port  $k$  when ports  $i, j, \dots$  are infinite

valued (shorted capacitors and opened inductors). In the presence of inductors a similar line of argument can be applied, noting that the time constant  $\tau_k^{ijk\dots}$  associated with inductor  $L_k$  is simply the inductance divided by  $R_k^{ijk\dots}$  which is the resistance seen at port  $k$  with the reactive elements at ports  $i, j, \dots$  at their infinite values<sup>56</sup>. So the time constants in (4.87) and (4.88) will have one of the following forms depending on whether there is an inductor or a capacitor connected to port  $k$ . For capacitor,  $C_i$ :

$$\tau_i^{ijk\dots} = C_i R_i^{ijk\dots} \quad (4.90)$$

and for inductor,  $L_l$ :

$$\tau_l^{mn\dots} = \frac{L_l}{R_l^{mn\dots}} \quad (4.91)$$

Finally, the transfer constant,  $H^{ijk\dots}$ , is the transfer function evaluated with the reactances at ports  $i, j, k, \dots$  at their infinite values (shorted capacitors and opened inductors) and all others zero valued (opened capacitors and shorted inductors). It is noteworthy that (4.87) indicates that the poles of the transfer function are independent of the definition of input and output and are only characteristics of the network itself, while the zeros are not a global property of the circuit and depend on the definition of the input and output ports and variables, as evident from the presence of the term  $H^{ijk\dots}$  terms. This is consistent with our observation that poles are the roots of the determinant of the  $Y$  matrix, as discussed in Chapter 2.

Several observations are in order about this approach. First of all this approach is exact and allows one (with enough patience in the case of a large network) to determine the transfer function completely and exactly with less effort compared to writing nodal or mesh equations (KCL and KVL) should the need arise. But more importantly, unlike writing nodal or mesh equations, one does not need to carry the analysis to its end to be able to obtain useful information about the circuit. One could just do the zero-value time constant analysis to obtain  $b_1$  and hence obtain an estimate of the dominant pole if such a dominant pole exist. However, if additional information about higher order poles and zeros are needed one can carry the analysis through enough steps to obtain the results to the desired level of accuracy. Also, the analysis is equally applicable to real and complex poles and zeros. Once mastered, this analysis method provides a fast and accurate means of determining transfer functions, as well as input and output impedances for general circuits.

The generalized time and transfer constants (TTC) approach has several important and useful corollaries that we will summarize next.

Online YouTube lecture on behavior of basic amplifier stages:

**High Frequency behavior of basic amplifier stages (CS, CE, CD, CC, CG, CB) using TTC**

<sup>56</sup> Equations (4.83) can be generalized noting the invariance of the  $\beta_l^{ijk\dots}$  to a rotation of the indexes to produce

$$R_i^0 R_j^i R_k^{ij} \dots R_m^{ijk\dots} = R_j^0 R_k^j R_l^{jk} \dots R_i^{jkl\dots} \quad (4.89)$$

### 4.6.1 Several Corollaries of the TTC Method

We have alluded to the first important corollary of the TTC methods in the previous section several times already. It is that the number of the poles in a system is equal to the maximum number of independent initial conditions we can set for energy-storing elements, or as is often stated the number of independent energy-storing elements. Let us first look at a relatively simple and useful example of what happens when some of the elements are not independent from the others.

#### Example 4.6.1 (Common-Emitter with Capacitive Load: Full Transfer Function)

Consider the common emitter stage of Example 4.3.3, shown in Figure 4.31. That circuit has three capacitors, but in fact we can only set two independent initial conditions. To see why, let us assume we set the voltage across  $C_\pi$  to  $V_\pi$  and the voltage of  $C_L$  to  $V_L$ . In that case the voltage across  $C_\mu$  is already determined to be  $V_\pi - V_L$ . Therefore we have only two independent degrees of freedom.

We have already determined the coefficient  $b_1$  in (4.43). Now let us quickly determine  $b_2$  using (4.82). To do so, we determine three time constants by short-circuiting the associated element with the superscript and looking at the impedance seen by the elements designated by the subscripts. Unlike ZVTs all of which we needed, there are six such combinations of these time constants ( $\tau_\mu^\pi$ ,  $\tau_\mu^L$ ,  $\tau_\pi^L$ ,  $\tau_\pi^\mu$ ,  $\tau_L^\mu$ , and  $\tau_L^\pi$ ), out of which we have to pick any three so that we cover any two-way combination once and only once to be coupled with the ZVTs. There are many combinations, but since we notice that  $\tau_\mu^0$  was the most complicated ZVT, we try to pick the ones that avoid that one to make our calculation more straightforward, i.e.,

$$\begin{aligned}\tau_\mu^\pi &= C_\mu R_2 \\ \tau_\pi^L &= C_\pi(r_\pi \parallel R_1) \\ \tau_\mu^L &= C_\mu(r_\pi \parallel R_1)\end{aligned}$$

that are calculated using the circuits shown in Figure 4.67. These combined with the ZVTs calculated in (4.43) produce:

$$\begin{aligned}b_2 &= \sum_i^{1 \leq i < j \leq 3} \sum_j \tau_i^0 \tau_j^i \\ &= \tau_L^0 \tau_\pi^L + \tau_\pi^0 \tau_\mu^\pi + \tau_L^0 \tau_\mu^L \\ &= C_L R_2 \cdot C_\pi(r_\pi \parallel R_1) + C_\pi(r_\pi \parallel R_1) \cdot C_\mu R_2 + C_L R_2 \cdot C_\mu(r_\pi \parallel R_1) \\ &= (r_\pi \parallel R_1) R_2 \cdot (C_\pi C_\mu + C_\pi C_L + C_\mu C_L) = R_{left} R_{right} \cdot C_\Delta^2\end{aligned}$$

where we define

$$C_\Delta^2 \equiv C_\pi C_\mu + C_\pi C_L + C_\mu C_L \quad (4.92)$$

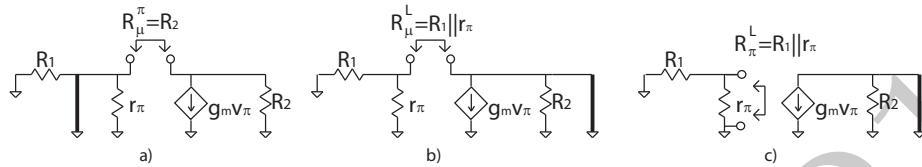


Figure 4.67: The equivalent circuit used to calculate for the common-emitter stage of Figure 4.31: a)  $\tau_{\mu}^{\pi}$ , b)  $\tau_{\mu}^L$ , c)  $\tau_{\pi}^L$ .

The next step is to determine  $b_3$ . From (4.87) we see that with three energy-storing elements,  $b_3 = \tau_1^0 \tau_2^1 \tau_3^{12}$ . Now we notice that  $\tau_3^{12}$  involves two infinite value subscripts and hence for this example, is proportional to the resistance seen by one of the capacitors, when the other two are shorted. In this case, since we have a loop, if two capacitors are shorted the third one is guaranteed to see a zero impedance and hence  $\tau_3^{12}$  and hence  $b_3$  are zero for any combination of the indexes. Thus the system is only second order with two poles.

The (4.88) results in the same RHP zero as in Example 4.2.5, since the only non-zero H coefficients are  $H^0$  and  $H^{\mu}$  and they have the same ratio as before, leading to  $z = g_m / C_{\mu}$ .

As can be seen from this example, the capacitive loop of three capacitors reduces the number of the independent energy-storing elements by one and thus lowers the order of the denominator's order (the number of the poles) by one.

In general, it is easy to verify that in a circuit with  $N + 1$  nodes (counting the ground as a node) with any number of capacitors between any number of nodes, we can only define  $N$  independent initial conditions and hence can have a maximum of  $N$  poles<sup>57</sup>. This is while in this case, we can have up to  $N(N+1)/2$  capacitors among various nodes.

A node with only inductors attached to it plays the same role as a loop of capacitors, since if we determine the initial current of all but one inductor, we have no freedom in choosing the current of the last one since it can be determined using KCL at that node.

Online YouTube lecture on behavior of basic amplifier stages:

#### Uncoupled time-constants, independent poles and zeros, observable dynamics

The second important corollary of TTC relates to *uncoupled* poles of the circuits. As we mentioned earlier, in general, there is no one-to-one correspondence between the poles' characteristic times ( $-1/p_i$ ) and the zero-valued time constants,  $\tau_i^0$ , as there is not even an equal number of them in general. However, an important exception is when a time-constant is decoupled from the other ones, namely, when its value does not change for any combination of shorting and opening of other energy-storing elements. If we assume that the time-constant

<sup>57</sup>We can see this from the fact that  $Y$  matrix  $N \times N$  and is of the order of  $N$  in  $s$ .

associated with the  $n$ th energy-storing element is decoupled from the rest, in the language of TTC this means that if we have

$$\tau_n^0 = \tau_n^i = \tau_n^{ij} = \dots = \tau_n^{ijk\dots m} \quad (4.93)$$

then the term  $(1 + \tau_n^0 s)$  can be factored out of the denominator and the pole associated with it is simply,  $p_n = -1/\tau_n^0$ .

This can be proved rather easily by expressing the denominator of the transfer functions as:

$$\begin{aligned} D(s) &= 1 + b_1 s + b_2 s^2 + \dots \\ &= 1 + s \sum_{i=1}^n \tau_i^0 + s^2 \sum_i \sum_{j \neq i}^n \tau_i^0 \tau_j^i + \dots \\ &= 1 + s(\tau_n^0 + \sum_{i=1}^{n-1} \tau_i^0) + s^2 (\sum_{i=1}^{n-1} \tau_i^0 \tau_n^i + \sum_i \sum_{j \neq i}^{n-1} \tau_i^0 \tau_j^i) + \dots \\ &= (1 + s\tau_n^0) + (1 + s\tau_n^0)s \sum_{i=1}^{n-1} \tau_i^0 + (1 + s\tau_n^0)s^2 \sum_i \sum_{j \neq i}^{n-1} \tau_i^0 \tau_j^i + \dots \\ &= (1 + s\tau_n^0) \left[ 1 + s \sum_{i=1}^{n-1} \tau_i^0 + s^2 \sum_i \sum_{j \neq i}^{n-1} \tau_i^0 \tau_j^i \dots \right] \end{aligned} \quad (4.94)$$

where the term in the bracket is of order of  $s^{n-1}$ .

This concept can be generalized to a group or groups of time constants that can be uncoupled from the rest of the time constants but internally coupled. An example is a multi-stage amplifier, with no interstage capacitors, where speaking the time constants within each stage may be coupled and cannot be factored into products of first order terms, however, it is possible to factor the numerator and denominator into product of lower order polynomials each associated with one set of externally uncoupled yet internally coupled set of time constants internal to each stage. This can be viewed as a partitioning of time constants into these mutually uncoupled subsets.

Yet another corollary of the TTC pertains to the number of zeros. We saw earlier that the number of poles is equal to the number of independent energy storing elements in the circuit. The number of zeros is determined by the order of the numerator polynomial, which is in turn determined by the highest order non-zero  $H^{ijk\dots}$  in (4.88). In other words, the number of zeros in the circuit is equal to the maximum number of energy-storing elements that can be *simultaneously* infinite-valued while producing a non-zero transfer constant  $H^{ijk\dots}$ . This was we can easily determine how many zero there are in the transfer function of the system by inspection without having to write any equations.

As a possible example of this corollary, consider the dual path system of Example 4.4.4 shown in Figure 4.47. There are two capacitors,  $C_1$  and  $C_2$ , shorting of either one results in a non-zero transfer constant. Nonetheless, there

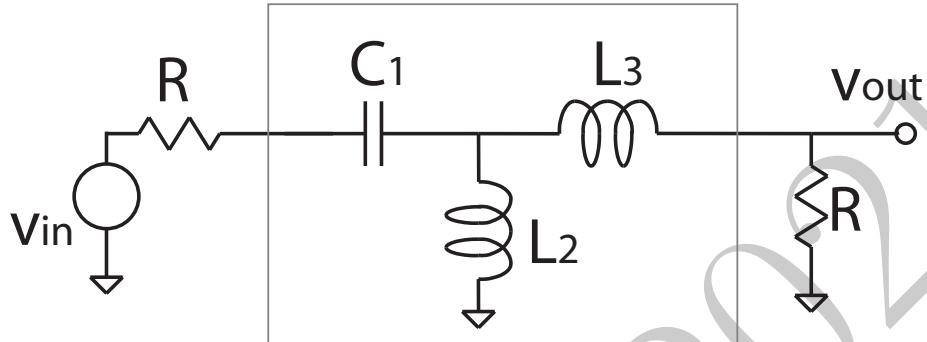


Figure 4.68: A third-order reactive bandpass filter.

is only one zero in the transfer function, since simultaneous shorting (infinite-valuing) of both results in a zero transfer constant ( $H^{12} = 0$ ).

As another case, consider the bandpass stage of Example 4.5.7 shown in Figure 4.65. In this case, taking the capacitor or the inductor to their infinite values one at a time results in a non-zero transfer constant only for the capacitor, since an infinite valued inductor (open) results in a zero transfer constant ( $H^L = 0$ ). However, we should not prematurely (and incorrectly) conclude that there is only one zero in the transfer function since the simultaneously infinite valued inductor and capacitor result in a *non-zero* transfer constant,  $H^{LC}$ , in this case. Therefore, there are two zeros in the system<sup>58</sup>.

Finally here is an example of a system with a three zeros and two poles:

**Example 4.6.2 Reactive Bandpass Filter** In this example we apply the approach to determine the exact transfer function of the reactive bandpass network of Figure 4.68. The time constants are:

$$\begin{aligned}\tau_1^0 &= RC_1 & \tau_2^0 &= L_2/R & \tau_3^0 &= L_3/R \\ \tau_2^1 &= 2L_2/R & \tau_3^1 &= L_2/R & \tau_3^2 &= 0 \\ \tau_3^{12} &= L_3/2R\end{aligned}$$

All transfer constants are zero with the exception of

$$H^{12} = \frac{1}{2}$$

which immediately results in the following transfer function:

$$H(s) = \frac{L_2 C_1 s^2}{1 + (RC_1 + \frac{L_2 + L_3}{R})s + (2L_2 C_1 + L_3 C_1)s^2 + \frac{L_2 L_3 C_1}{R}s^3}$$

demonstrating the ease of application of the method to a passive lossless reactive network.

<sup>58</sup>Note that in this example,  $a_1$  is zero while  $a_2$  is not, thus the zeros for a conjugate imaginary pair on the  $j\omega$ -axis.

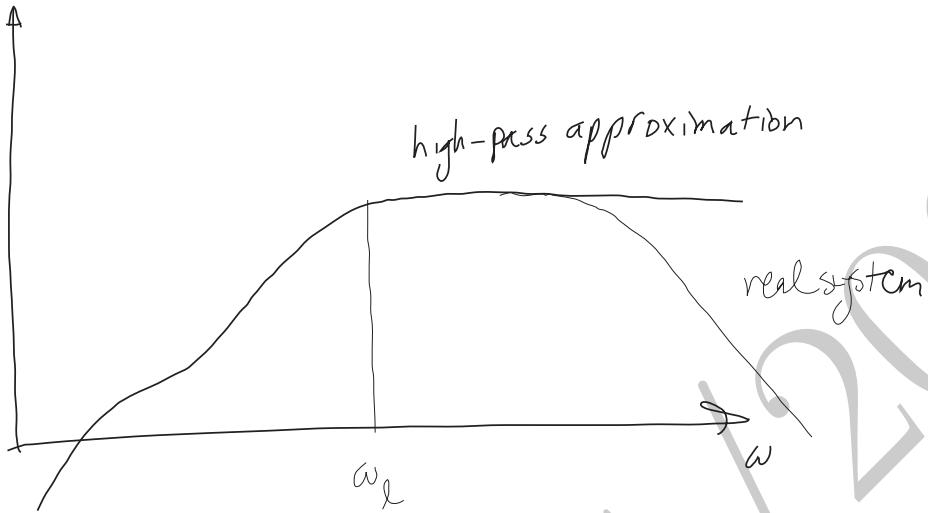


Figure 4.69: The transfer function of an  $N$ th order high-pass system and the original bandpass system it approximates.

#### 4.6.2 Infinite Value Time Constants

Online YouTube lecture:

[Low Cut-off Frequency Estimation, Infinite value time-constants \(IVT\)](#)

We saw earlier in (4.4) that the transfer function of a bandpass system can be factored into the part responsible for the low-frequency behavior in terms of inverse poles and zeros, which results in a high pass response and a part responsible for the high-frequency behavior in terms of conventional poles and zeroes that form a low pass response. So far most of our discussions have been on the high frequency behavior of a low-pass system. However, we can apply a special case of TTCs called the infinite value time-constant (IVT) approach to determine the low-frequency behavior.

It applies to a *high-pass* system to determine its *low -3dB* frequency,  $\omega_l$ . Consider the high-pass transfer function shown in Figure 4.69. To have a unity response at high frequencies, the numerator should be of the same order as the denominator. One such case is:

$$H(s) = \frac{a_n s^n}{1 + b_1 s + b_2 s^2 + \dots + b_n s^n} = \frac{a_{mid}}{1 + \frac{b_{n-1}}{b_n s} + \dots + \frac{1}{b_n s^n}} \quad (4.95)$$

where  $a_{mid} = a_n/b_n$  is the gain at very high frequencies. The most dominant term affecting  $\omega_l$  is  $b_{n-1}/b_n$ .

Let us first look at the special case of a second order high-pass system with

the transfer function:

$$H(s) = \frac{a_2 s^2}{1 + b_1 s + b_2 s^2} = \frac{a_{mid}}{1 + \frac{b_1}{b_2 s} + \frac{1}{b_2 s^2}} \quad (4.96)$$

whose low cut-off frequency,  $\omega_l$  can be approximated as

$$\omega_l \approx \frac{b_1}{b_2} = \frac{\tau_1^0 + \tau_2^0}{\tau_1^0 \tau_2^1} = \frac{\tau_1^0}{\tau_1^0 \tau_2^1} + \frac{\tau_2^0}{\tau_2^0 \tau_1^2} = \frac{1}{\tau_1^1} + \frac{1}{\tau_2^2} \quad (4.97)$$

where we have used the equality  $\tau_1^0 \tau_2^1 = \tau_2^0 \tau_1^2$  from (4.64). As we can see, the low cut-off frequency can be determined by the sum of the reciprocals of the time-constants associated with each element, when the other one is infinite valued.

In general for an  $n$ th order high-pass system, we can approximate  $\omega_l$  with the first order *inverse pole*, to obtain:

$$\begin{aligned} \omega_l \approx \frac{b_{n-1}}{b_n} &= \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n \sum_{k=j+1}^n \dots \sum_{m=n+1}^N \tau_i^0 \tau_j^i \tau_k^{ij} \dots}{\tau_i^0 \tau_j^i \tau_k^{ij} \dots \tau_n^{ij\dots m}} \\ &= \frac{1}{\tau_1^{23\dots n}} + \frac{1}{\tau_2^{13\dots n}} + \dots + \frac{1}{\tau_n^{12\dots (n-1)}} \end{aligned} \quad (4.98)$$

where we have used the rotational symmetry discussed in the footnote on Page 204. The time constant,  $\tau_i^{12\dots(i-1)(i+1)\dots n}$ , which we will show as,  $\tau_i^\infty$ , is the time constant for the  $i$ th element with all other ports at infinite values and hence an *infinite value time constant IVT*<sup>59</sup>.

This can be summarized as:

$$\boxed{\omega_l \approx \frac{b_{n-1}}{b_n} = \sum_{i=1}^N \frac{1}{\tau_i^\infty}} \quad (4.99)$$

where,

$$\tau_i^\infty = C_i R_i^\infty \quad (4.100)$$

for capacitor,  $C_i$ , and

$$\tau_l^\infty = \frac{L_l}{R_l^\infty} \quad (4.101)$$

for inductor  $L_l$ . Resistance  $R_i^\infty$  is the resistance seen looking into port  $i$  when the capacitors and inductors at *all other* ports are at their infinite values (shorted capacitors and opened inductors).

<sup>59</sup>When the energy-storing elements are capacitors only, this method is often referred to as the method of *short-circuit time constants*. Of course, the term infinite value time-constant is advantageous because it applies to both capacitors and inductors.

**Example 4.6.3 AC-Coupled Common-Emitter:** Consider the bandpass common-emitter stage of Example 4.3.8 shown in Figure 4.41a, with the bandpass transfer function shown in Figure 4.42. There is a coupling capacitor,  $C_c$ , between the input source resistance and the input of the stage.

The ZVTs can be easily calculated by computing the resistors seen by each of the two capacitors when the other one is zero-valued, i.e.,

$$\begin{aligned} R_\pi^0 &= r_\pi \\ R_c^0 &= r_\pi + R_1 \end{aligned}$$

and hence

$$b_1 = \tau_\pi^0 + \tau_c^0 = R_\pi^0 C_\pi + R_c^0 C_c = r_\pi C_\pi + (r_\pi + R_1) C_c$$

Since  $\tau_\pi^c = (R_1 \parallel r_\pi) C_\pi$ , we have

$$b_2 = \tau_c^0 \tau_\pi^c = r_\pi R_1 C_\pi C_c$$

Obviously,  $a_0$  is zero. For the rest of the  $H$  coefficients we have:

$$H^c = -g_m R_2 \cdot \frac{r_\pi}{r_\pi + R_1} = a_{mid}$$

$$H^\pi = 0$$

$$H^{c\pi} = 0$$

where  $a_{mid}$  is the mid-band gain of the amplifier. Hence we have

$$\begin{aligned} a_1 &= \tau_c^0 H^c = \tau_c^0 \cdot a_{mid} = -g_m R_2 r_\pi C_c \\ a_2 &= 0 \end{aligned}$$

For the coupling capacitor to be useful we should have  $C_c \gg C_\pi$  hence the transfer function can be written as

$$\begin{aligned} H(s) &= \frac{a_1 s}{1 + b_1 s + b_2 s^2} = \frac{a_{mid} \tau_c^0 s}{1 + (\tau_c^0 + \tau_\pi^0)s + \tau_c^0 \tau_\pi^0 s^2} \\ &\approx \frac{a_{mid} \tau_c^0 s}{(1 + \tau_c^0 s)(1 + \tau_\pi^0 s)} \approx \frac{1}{1 + \frac{1}{\tau_c^0 s}} \cdot a_{mid} \cdot \frac{1}{1 + \tau_\pi^0 s} \end{aligned} \quad (4.102)$$

which is written as an inverse pole at  $\omega_l \approx 1/\tau_c^0 = 1/(r_\pi + R_1)C_c$  (i.e., a zero at the origin and a pole at  $\omega_l$ ) on the left describing the low frequency behavior and a pole at  $\omega_h \approx -1/\tau_\pi^0$  describing the high-frequency response on the right hand side.

As you can see, applying the TTC to this example, the results of ZVT and IVT can be seen easily. As for the ZVT, the procedure described in subsection 4.3.1, where all the elements whose infinite values improve the transfer function, are set to their infinite values, would correspond to calculating the time constant associated with  $C_\pi$  when  $C_c$  was short circuited. This is consistence with the high

frequency bandwidth,  $\omega_h$  being determined by the time constant,  $\tau_\pi^c$  instead of  $\tau_\pi^0$ , as in the high frequency equivalent circuit shown in Figure 4.43 of Example 4.3.8. Also you can see that the inverse pole frequency (and hence  $\omega_l$  is determined by the time constant seen by  $C_c$  when  $C_\pi$  is infinite valued.

Finally, for an example of how the design progression of EXXX is implemented in discrete and integrated setting and how different they finally look, watch the following online YouTube lecture (it also includes and example of IVT application):

[\*\*Discrete vs. Integrated Variations, Low Cut-Off Estimation with IVT example\)\*\*](#)

A discussion of broadband amplifier design concept can be found in the following online YouTube lecture (it also includes and example of IVT application):

[\*\*Broadband amplifiers, gain-bandwidth product, optimum gain per stage\)\*\*](#)

# Chapter 5

## Feedback Viewpoint

Although all of the circuits with feedback can be directly analyzed by direct application of KVL and KCL (e.g., nodal analysis), the feedback viewpoint is much more helpful to design as it offers a causal relationship between the amount and nature of the feedback applied and the general properties of the loop. In particular, the loop gain is a key parameter in determining the stability of a system with a feedback loop.

### 5.1 Infinite Forward Gain

Online YouTube lecture:

[\*\*Feedback: Asymptotic Transfer Function, 1st Order Analysis, Rapid design\*\*](#)

In this section, we assume that we can scale the gain in the forward path by introducing a scaling factor  $k$  that can be varied. By making this factor arbitrarily large ( $k \rightarrow \infty$ ) we look at the effect of the infinite forward gain on the feedback network asymptotically. This assumption allows us to first see the desired behavior of the circuit with feedback without having to deal with the circuit non-idealities. We will see how various non-idealities can be accounted for in the subsequent sections.

Consider an amplifier with two differential inputs  $u_{i+}$  and  $u_{i-}$  and a single output  $u_o$ . The input and output variables could be currents and/or voltages, however the two inputs must be of the same kind, i.e., either both of them are currents or voltages. The amplifier has a gain of  $A$ , meaning the output is related to the input via

$$u_o = A \cdot (u_{i+} - u_{i-}) \quad (5.1)$$

Scaling the forward path gain  $A$  by a factor  $k$  and letting  $k \rightarrow \infty$ , allows us to capture the most important part of the feedback in a circuit.

Now, let us first assume that we feed a fraction of the output variable,  $u_o$  back to the negative input  $u_{i-}$  using a *scaler* feedback network,  $f$ , as shown in Figure 5.1. By scaler we mean that the output of the feedback network is a

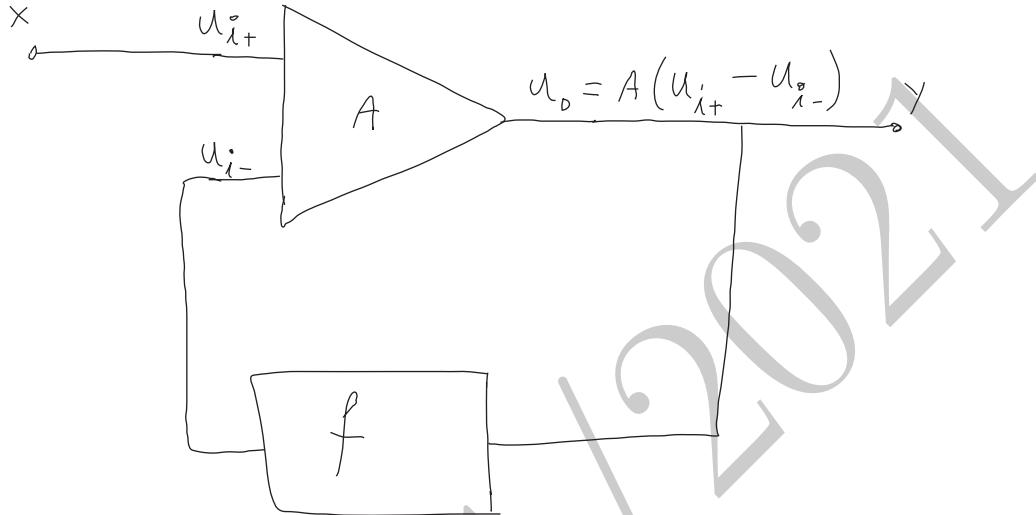


Figure 5.1: An ideal amplifier with a differential input and single-ended output with scalar feedback factor,  $f$ . The input and output variables to the amplifier and the feedback network could be voltages or currents.

linearly scaled version of its input, i.e., its output is simply its input multiplied by a constant,  $f$ . We assume that the feedback block,  $f$ , has the appropriate input and output variables compatible with  $u_o$  and  $u_i$ , respectively<sup>1</sup>. We also assume for the time being that both the amplifier and the feedback blocks are unilateral<sup>2</sup>, meaning that the signal flows only from their input (on the left for the amplifier and on the right for the feedback network in Figure 5.1) to their outputs with no reverse signal gain from the outputs to the inputs.

The transfer function between the overall system input  $x$  and its output  $y$  can be easily calculated by noting that the output is related to the input through

$$y = A(u_{i+} - u_{i-}) = A(x - fy)$$

which can be solved to obtain

$$H \equiv \frac{y}{x} = \frac{A}{1 + Af} \quad (5.2)$$

If we scale the forward path gain,  $A$ , by a factor of  $k$  and make the new forward path gain ( $kA$ ) go to infinity by letting  $k \rightarrow \infty$  (Figure 5.2), then 5.2 reduces to:

$$H_\infty \equiv H|_{k \rightarrow \infty} = \frac{kA}{1 + kA}|_{k \rightarrow \infty} = \frac{1}{f} \quad (5.3)$$

<sup>1</sup>For example, if the amplifier's output is a voltage, the feedback network's input is a voltage too.

<sup>2</sup>We will see how to account for the bilateral case in subsection 5.4.2.

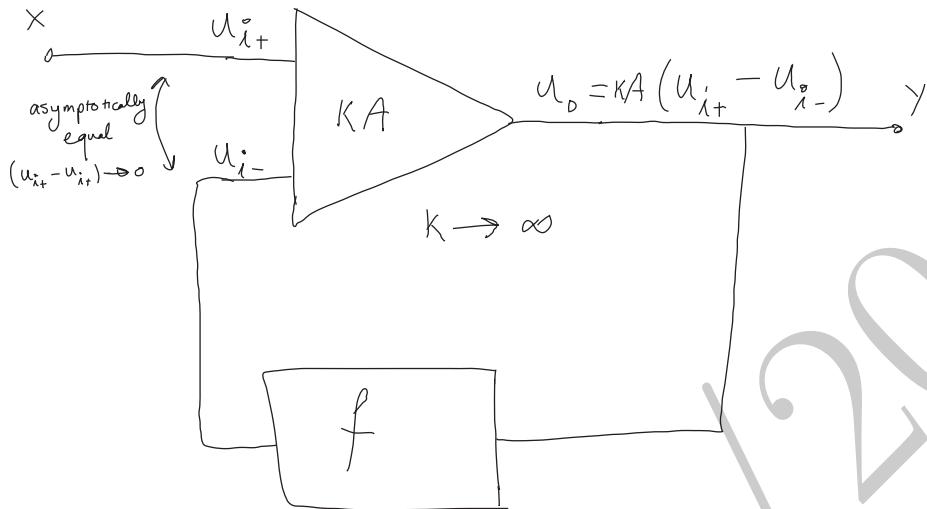


Figure 5.2: The feedback configuration of Figure 5.1 where the forward amplifier gain is scaled by a factor of  $k$ . Making  $k$  very large ( $k \rightarrow \infty$ ), the two inputs must be asymptotically equal to have a finite output.

We define the transfer function under these conditions as the *asymptotic transfer function* with reference to  $A$ .

This is an important and useful result as it indicates that if the gain of the forward path is made large enough, the transfer function of the closed-loop system is *solely* determined by the feedback network and is independent of the forward path characteristics. Also it is an important observation that the closed-loop system gain is the reciprocal of the feedback factor  $f$  (i.e.,  $1/f$ ). We will see that the inversion of the feedback function holds in a broad range of feedback network types.

### Asymptotic Equality Principle

There is another way to arrive at the result of (5.3) that is more useful in circuit design. Assuming negative feedback, we should have a finite steady-state output,  $y$ , for any finite input  $x$  to the system in Figure 5.1. However, since the amplifier forward gain is made infinite (via  $k \rightarrow \infty$ ), the only way for its output to have a finite value is for the difference between its two inputs to be infinitesimal<sup>3</sup>, namely,  $k \rightarrow \infty$  results in  $(u_{i+} - u_{i-}) \rightarrow 0$  and the transfer function will approach the asymptotic transfer function  $H_\infty \equiv H|_{k \rightarrow \infty}$ . We will refer to this as the *asymptotic equality principle*<sup>4</sup>.

<sup>3</sup>In practice the higher the forward gain, the “more equal” the two inputs will be to quote George Orwell.

<sup>4</sup>A special form of this principle is commonly referred to as the *virtual ground* principle when dealing with voltage amplifiers such as operational amplifiers.

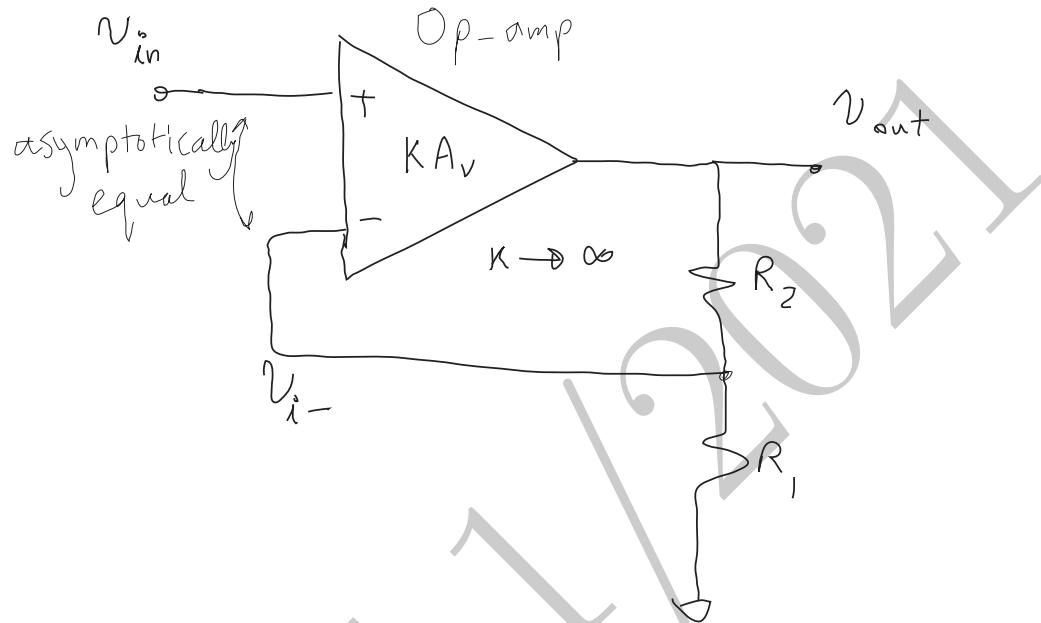


Figure 5.3: An operational amplifier connected as a non-inverting voltage amplifier. Series-shunt (voltage-voltage) configuration.

In practice, it is often more straightforward to use this approach to determine  $H_\infty$  which can be then used to calculate  $f$  using (5.3), as opposed to trying to determine  $f$  to calculate  $H_\infty$ , as we will see in several examples. This is mainly due to the fact that except for the most straightforward of cases, in practical circuits it may not be clear where the forward path ends and where the feedback network begins. In most transistor circuits, the feedback network is also part of the biasing or load network of the main amplifier and cannot be easily removed or isolated without affecting the forward path.

We will see in the following examples how the *asymptotic equality principle* can be applied and the consistency of the results obtained using that principle with the expected feedback factor,  $f$ .

**Example 5.1.1 (Operational Amplifier in Non-Inverting Configuration)**  
An operation amplifier (Op-Amp) is an amplifier where the input and output variables are all voltages and hence the input impedances is large and the output impedance is small.

Let us consider the non-inverting voltage amplifier configuration shown in Figure 5.3 made with an ideal op-amp (infinite gain and input impedance, zero output impedance). It is obvious in this case, that the input and the output variables are voltages, shown with  $v$ 's instead of  $u$ 's. Now we apply the asymptotic equality principle to this amplifier. The two inputs must have the same voltage

for infinite forward gain (the only way we can have a finite output). Therefore the output voltage must adjust itself in such a way that the voltage across  $R_1$  is equal to  $v_{in}$  by negative feedback. However, the voltage at the negative terminal is simply  $v_{out}$  divided down by the resistive voltage divider, resulting in the following asymptotic transfer function with respect to  $A$ ,

$$H_\infty \equiv \frac{v_{out}}{v_{in}}|_{k \rightarrow \infty} = 1 + \frac{R_2}{R_1}$$

The above result can be used to determine  $f$  (should the need arise) using (5.3) to be

$$f = \frac{1}{H_\infty} = \frac{R_1}{R_1 + R_2}$$

We notice that in this case, this is simply the voltage divider ratio of the voltage feedback network formed by  $R_1$  and  $R_2$  in Figure 5.3 for voltage input and outputs and no loading by the input terminal of the op-amp.

Feedback configurations of this nature where both the sensed and returned quantities are voltages are referred to as *voltage-voltage* or *series-shunt* configuration. The reason for the later terminology is that to sense a voltage the meter is applied in *shunt* (parallel) and that to subtract a voltage from another, the voltage sources should be placed in *series*.

Note that order of the names is reverse in the two naming conventions: in the sense-return naming convention (e.g., voltage-voltage) the first quantity is the sensed quantity (often the *output*) and the second one is the returned quantity (often the *input*). In the input-output naming convention, the input configuration (i.e., shunt or series) appears first and the the output configuration.

Here is another example of *series-shunt* configuration using transistors

**Example 5.1.2 (MOS series-shunt feedback amplifier)** The MOS amplifier of Figure 5.4a consists of two MOS transistors ( $M_1$  is an NMOS and  $M_2$  is a PMOS). In addition to providing the dc bias current to  $M_1$  and presenting a resistive load to  $M_2$ , the resistive network consisting of  $R_1$  and  $R_2$  senses the output voltage and returns a voltage proportional to  $v_{out}$  in series with the source of the transistor  $M_1$  which is subtracted from the input voltage,  $v_{in}$ , to produce the small-signal gate-source voltage,  $v_{gs}$ .

Now we need to let the gain of the forward amplifier comprising  $M_1$  and  $M_2$  to approach infinity without changing the feedback network. This can be done, for instance, by scaling the  $\pi$ - or T-model dependent current sources of  $M_1$  or  $M_2$  with a factor  $k$  and letting it go to infinity<sup>5</sup>. Note that we do not explicitly show the current gain of the T-model of transistor  $M_1$  since it is normally unity. For  $k \rightarrow \infty$ , we notice that a finite output ac voltage implies a finite ac voltage at the gate of  $M_2$  (since  $g_{m2}$  is finite). This in turn implies a finite ac drain current,  $i_{d1}$  for  $M_1$ . However, the only way to obtain a finite  $i_{d1} = kv_{gs1}/r_{m1}$  with an infinite

<sup>5</sup>It can even be done by scaling  $R_3$ , though this is a bit less convenient in calculation of the return ratio discussed in Section 5.2.

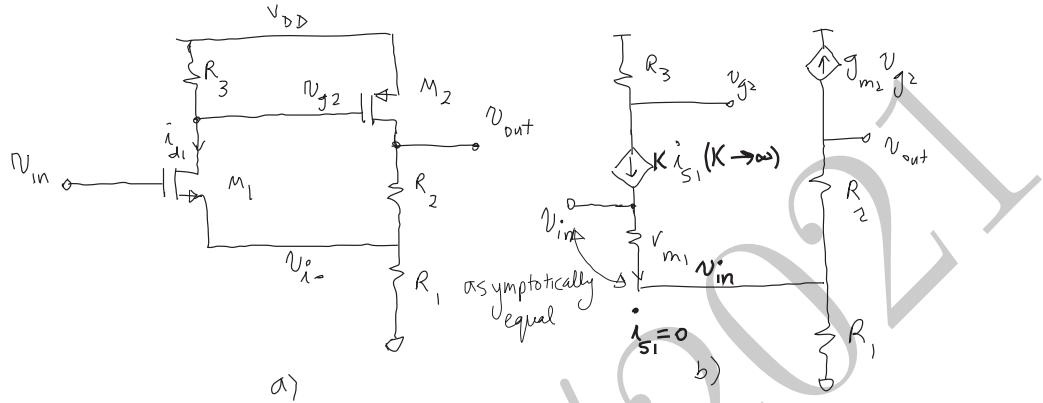


Figure 5.4: a) A MOS series-shunt (voltage-voltage) feedback amplifier, b) its small-signal equivalent where a T-model (with  $\alpha = 1$ ) is used for  $M_1$  and a  $\pi$ -model for  $M_2$ .

$k$  is for the small-signal source current and thereby the small-signal gate-source voltage of  $M_1$  to be zero ( $i_{s1} = 0$  and  $v_{gs1} = 0$ ). This means  $v_{in} = v_{i-}$ , i.e., the negative feedback with infinite forward gain forces an asymptotic equality (a virtual short) between  $v_{in}$  and  $v_{i-}$ . This simply means that  $v_{in}$  and  $v_{out}$  have to be related through the resistive divider formed by  $R_1$  and  $R_2$ , namely,

$$v_{in} = v_{i-} = v_{out} \cdot \frac{R_1}{R_1 + R_2}$$

which translates to an asymptotic transfer function with respect to the dependent current source of  $M_1$  in the T-model,

$$H_\infty \equiv \frac{v_{out}}{v_{in}}|_{k \rightarrow \infty} = 1 + \frac{R_2}{R_1} \quad (5.4)$$

This expression can be used to determine  $f$  using (5.3) to be

$$f = \frac{1}{H_\infty} = \frac{R_1}{R_1 + R_2}$$

which is the voltage divider ratio of the resistive feedback network in this case.

In practice, the asymptotic transfer function can serve as an approximation of the transfer function. We will see in Section 5.2 how much the transfer function deviates from the asymptotic transfer function, ( $H_\infty$ ), due to the finite value of the forward gain compared to the idealized result.  $H_\infty$  is the parameter we design to. In a design exercise, we initially choose the values of the components so that  $H_\infty$  is close to our desired transfer function.

Here is another transistor amplifier example:

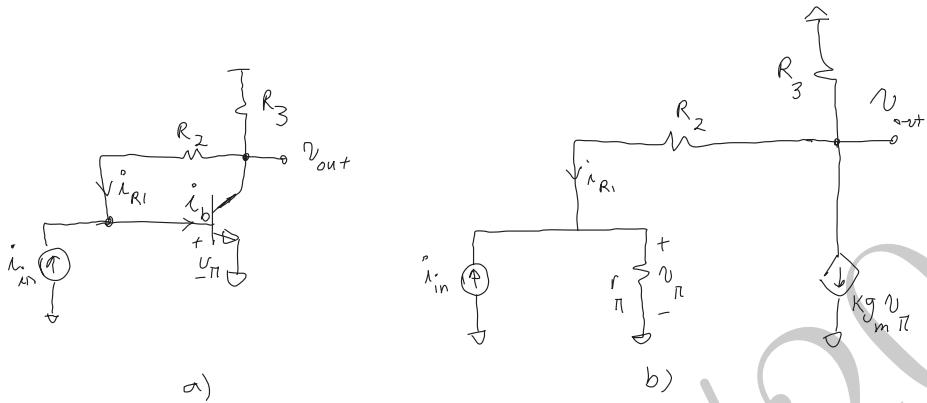


Figure 5.5: a) A BJT shunt-shunt (voltage-current) feedback amplifier, and b) its small-signal equivalent circuit.

**Example 5.1.3 (BJT Shunt-Shunt Amplifier)** In the common-emitter amplifier of Figure 5.5, resistor  $R_2$  provides a feedback path from the output to the input. In this case, the input is a current and the resistor  $R_2$  returns a current to the input node which is proportional to the output voltage.

To calculate the infinite-forward-gain transfer function, we can scale the transistor's transconductance,  $g_m$ , by  $k$  and let  $k$  go to infinity. The only way that we can have a finite small-signal  $v_{out}$  with an infinite transconductance  $k g_m$  is for the small-signal base-emitter voltage to be zero ( $v_\pi = 0$ ). This simply implies that the small-signal base current (the current through the  $r_\pi$ ) is zero ( $i_b = 0$ ). Therefore, the current through  $R_2$  is asymptotically equal to the input current,  $i_{in}$ , i.e., for  $k \rightarrow \infty$ , we have,

$$i_{in} = -\frac{v_{out}}{R_2}$$

where we have used the fact that the small-signal voltage on the left side of  $R_2$  is zero for infinite forward gain ( $v_\pi = 0$ ). Thus, the asymptotic transfer function with respect to  $g_m$  is given by:

$$H_\infty = \frac{v_{out}}{i_{in}}|_{k \rightarrow \infty} = -R_2$$

Note that in this case, input variable  $u_{i+}$  is the input current  $i_{in}$ , and input variable  $u_{i-}$  is the current through the resistor,  $i_{R_2} = -v_{out}/R_2$ . This implies that  $f$  which is the ratio of  $u_{i-}$  to  $u_o$  is simply,  $f = -1/R_2$  which is consistent with the result for  $H_\infty$  via (5.3). Also note that  $(u_{i+} - u_{i-})$  is simply  $i_b$  which is zero for infinite forward gain.

Here is another example:

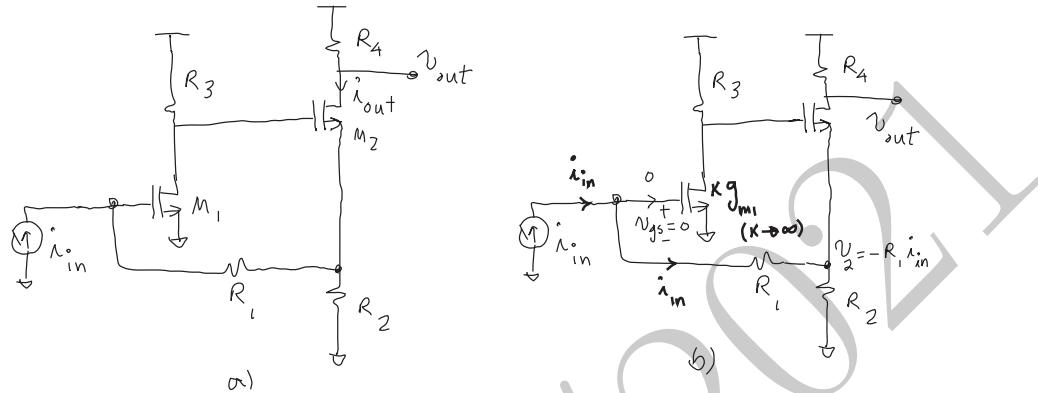


Figure 5.6: a) A shunt-series (current-current) feedback stage with a current input and voltage output, b) calculation of its asymptotic transfer function.

**Example 5.1.4 (MOS Shunt-Series Feedback Stage)** In this example we consider the shunt-series feedback stage of Figure 5.6a with a current input,  $i_{in}$ . Transistor  $M_1$  is connected in common-source configuration and  $M_2$  can be thought of as being a source-degenerated common source from output's perspective and a source follower from feedback point of view.

Let us calculate the asymptotic transfer function,  $H_\infty$ , with respect to the transconductance of  $M_1$  by scaling  $g_{m1}$  to  $kg_{m1}$  and let  $k$  go to infinity, as shown in Figure 5.6b. The only way to maintain a finite small-signal voltage at the drain of  $M_1$  when  $k \rightarrow \infty$  is for the small-signal gate-source voltage,  $v_{gs}$  to be zero. This also implies that the input current is zero. This is true even if the input impedance of  $M_1$  is not infinite, e.g., at high frequencies with an input capacitance or if it were a BJT. Hence the input current,  $i_{in}$  will flow directly through  $R_1$ , resulting in a voltage  $v_2 = -R_1 i_{in}$  at the source of  $M_2$ . The source current of  $M_2$ , which is the same as its drain current at low frequencies can be calculated by applying KCL at its source as the difference between  $i_{in}$  and the current through  $R_2$ :

$$i_{d2} = i_{s2} = \frac{v_2}{R_2} - i_{in} = -i_{in} \frac{R_1}{R_2} - i_{in} \quad (5.5)$$

which directly allows us to calculate the output voltage and hence the asymptotic transfer function:

$$H_\infty \equiv \frac{v_{out}}{i_{in}}|_{k \rightarrow \infty} = R_4 \cdot \left( 1 + \frac{R_1}{R_2} \right) \quad (5.6)$$

**Example 5.1.5 (MOS Shunt-Shunt Amplifier)** Consider the shunt-shunt feedback amplifier shown in Figure 5.7a, with an input current source  $i_{in}$  applied to the source of transistor  $M_2$  and the output voltage,  $v_{out}$  taken off of its drain.

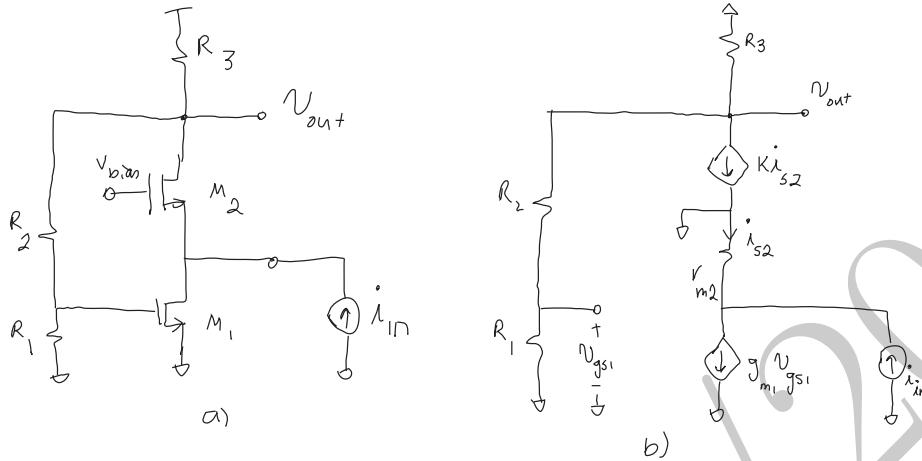


Figure 5.7: a) A MOSFET shunt-shunt (voltage-current) feedback circuit, and its small-signal equivalent model.

Let us calculate the asymptotic transfer function,  $H_\infty$ , defined as the ratio of the output voltage to the input current. It can be calculated with respect to various source. Let us calculate it with respect to the drain-gate dependent source in the T-model of  $M_2$ , shown in Figure 5.7b. The scaled dependent current source is stated as  $kis_2$  since  $\alpha_2 = 1$  for a MOSFET (equal drain and source currents).

To determine the asymptotic transfer function,  $H_\infty$ , with respect to this controlled source, we should let  $\alpha_2$  approach infinity. The only way to maintain a finite output voltage with an infinite  $k$  is for the small-signal source current,  $k$  to be zero. Since this is the current through the T-model gate-source resistance,  $1/g_{m2}$ , the small-signal gate-source voltage,  $v_{gs2}$ , must be zero too. In this case, asymptotic equality indicates that  $i_{in}$  must be equal to the small-signal drain current of  $M_1$ , which is  $g_{m1}v_{gs1}$ , i.e.,

$$i_{in} = i_{d1} = g_{m1}v_{gs1}$$

However, the small-signal voltage at the gate of  $M_1$  is related to  $v_{out}$  by the voltage divider ratio formed by  $R_1$  and  $R_2$ , namely,

$$v_{gs1} = v_{out} \cdot \frac{R_1}{R_1 + R_2}$$

Therefore, the asymptotic transfer function is given by,

$$H_\infty \equiv \frac{v_{out}}{i_{in}}|_{k \rightarrow \infty} = \frac{1}{g_{m1}} \left( 1 + \frac{R_2}{R_1} \right) \quad (5.7)$$

Here is another example with more transistors:

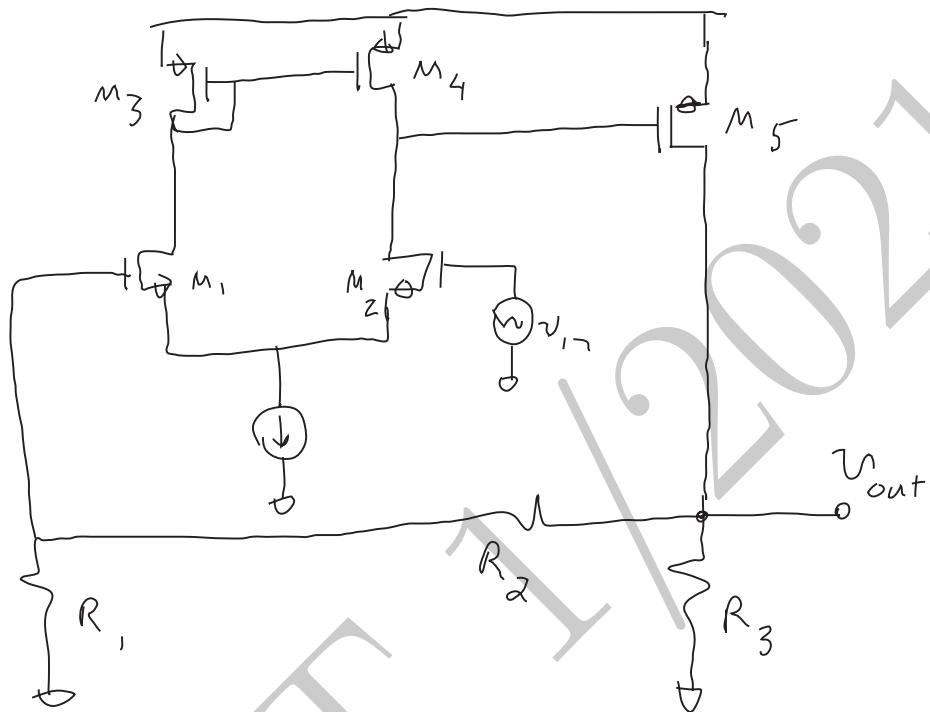


Figure 5.8: A two-stage MOS op-amp in series-shunt (voltage-voltage) feedback configuration.

**Example 5.1.6 (Two-stage differential MOS amplifier with series-shunt feedback)**

Consider the differential input amplifier of Figure 5.8. Transistor  $M_1$ - $M_4$  form a differential-to-single-ended amplifier with  $M_5$  operating as a second gain stage. The input voltage is applied to the input of  $M_2$ , while the feedback signal is returned to the gate of  $M_1$ . Calculating the asymptotic transfer function with respect to the T-model dependence current source of  $M_2$ , we can easily see that to have a finite output voltage, when  $k \rightarrow \infty$  the small-signal current through  $r_{m2}$  must be zero which in turn implies that the small-signal gate source voltage of  $M_2$  must be zero. Since the small signal impedance of the ideal tail current source to ground is infinity, we conclude that the small-signal current through  $r_{m1}$  which is the same as that through  $r_{m2}$  is also zero asymptotically. This means that the small-signal voltage at the gate of  $M_1$  is simply  $v_{in}$ . Therefore,

$$H_\infty \equiv \frac{v_{out}}{v_{in}}|_{k \rightarrow \infty} = 1 + \frac{R_2}{R_1}$$

This is not a surprising result, as this is essentially an op-amp in a non-inverting configuration.

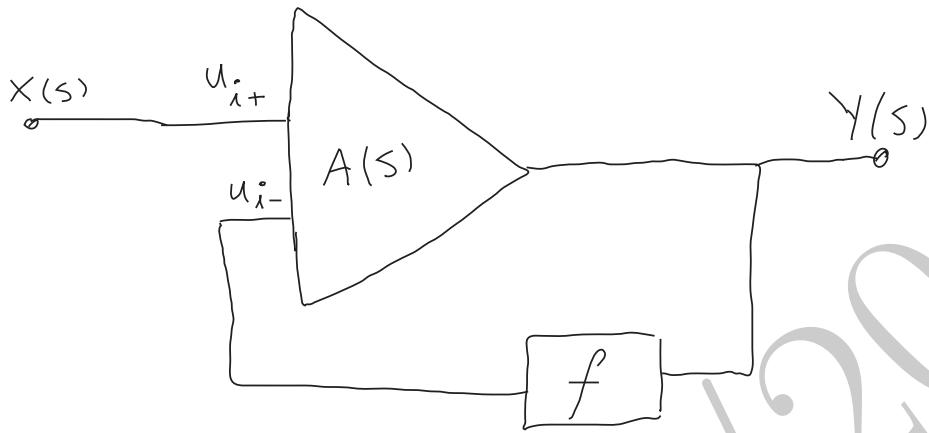


Figure 5.9: Amplifier with an LTI system in the feedback path.

### 5.1.1 Effect of Feedback on the Forward Path Dynamics

Online YouTube lecture:

[Feedback dynamics, forward and feedback path frequency effect, feedback sensitivity reduction](#)

In general, the forward path in the feedback configuration has a frequency dependent dynamics. This can be modeled by replacing the frequency independent forward path gain,  $A$ , with a transfer function,  $A(s)$  which has a low-frequency gain<sup>6</sup> of  $A_0$ , as illustrated in Figure 5.9.

If we simplistically assume that  $A(s) \rightarrow \infty$  irrespective of frequency, we simply obtain the same result as (5.3) without any additional insight. However, it is unrealistic to assume infinite forward gain at *all* frequencies. So let us look at two different asymptotic cases. First, at frequencies where  $A(s)f \gg 1$  (e.g., low frequencies), the transfer function of (5.2) simply reduces to  $H(s) = 1/f$ , as in (5.3). At the other extreme, when  $A(s)f \ll 1$ , the transfer function simply reduces to  $H(s) = A(s)$ , in other words, the open loop behavior of the system<sup>7</sup>. Although we have determined the behavior in these two limiting cases at this point, it is not clear how the transfer function behaves while transition from one region to the other. The behavior in the transitional region can vary a great

<sup>6</sup>In the case of a band-pass response,  $A_0$  would be the mid-band gain.

<sup>7</sup>In general, if the transfer function can be expressed as ratio of two polynomials in  $s$ , i.e.,  $A(s) = N(s)/D(s)$ , then the transfer function can be written as

$$H(s) = \frac{N(s)}{D(s) + f \cdot N(s)}$$

which shows that the zeros of the closed-loop transfer function are the same as the zeros of the forward path, while the poles are not constant and vary depending of the amount of feedback. The fact that the zeros remain stationary is understandable because they are the complex frequencies at which the forward transfer function from the input to the output is zero.

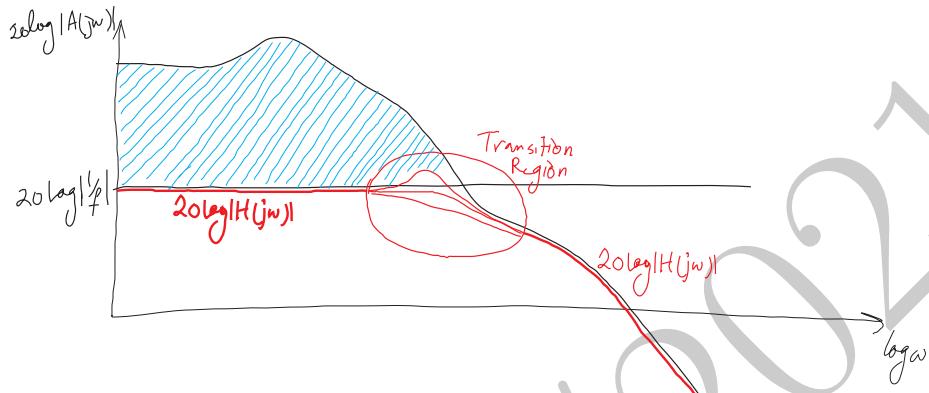


Figure 5.10: Amplifier with an LTI system in the feedback path.

deal and depends on certain parameters of the loop. It will be discussed in details in the Chapter 6. This is shown symbolically in a magnitude plot of  $H(j\omega)$  vs. frequency on a log-log scale (Bode Plot) in Figure 5.10.

An important observation at this point is that the application of feedback to this system has effectively increased the bandwidth of the closed-loop transfer function,  $H(s)$ . This is achieved at the cost of a lower closed-loop gain in frequency range where  $A(s)f \gg 1$ . So roughly speaking, the feedback loop has eliminated the extra open-loop gain in the hatched region of Figure 5.10. Although the bandwidth enhancement comes at the cost of gain reduction, it is a valuable tool to be able to trade the two in a simple fashion and hence achieve a high bandwidth in applications where a very large gain is not necessary. An added advantage of this arrangement is that as long as the feedback network is realized using a simple passive network with no significant frequency dependency, the in-band closed-loop transfer function will be flat, even in the presence of gain variations in the open-loop transfer function,  $A(s)$ , at the same frequencies.

The following example shows this principle in a simple first order system.

**Example 5.1.7 (First-Order Low-Pass Forward Path)** Let us assume that the forward pass transfer function,  $A(s)$ , is a first-order low-pass one, i.e.,

$$A(s) = \frac{A_0}{1 + \tau s} \quad (5.8)$$

where  $\omega_0 = 1/\tau$  is the 3dB bandwidth of the open-loop system. Now If we place this in a feedback configuration similar to Figure 5.9, the closed-loop transfer function will be given by

$$H(s) = \frac{A(s)}{1 + fA(s)} = \frac{A_0}{1 + A_0f + \tau s} = \frac{A_0}{1 + A_0f} \cdot \frac{1}{1 + \frac{\tau}{1+A_0f}s} \quad (5.9)$$

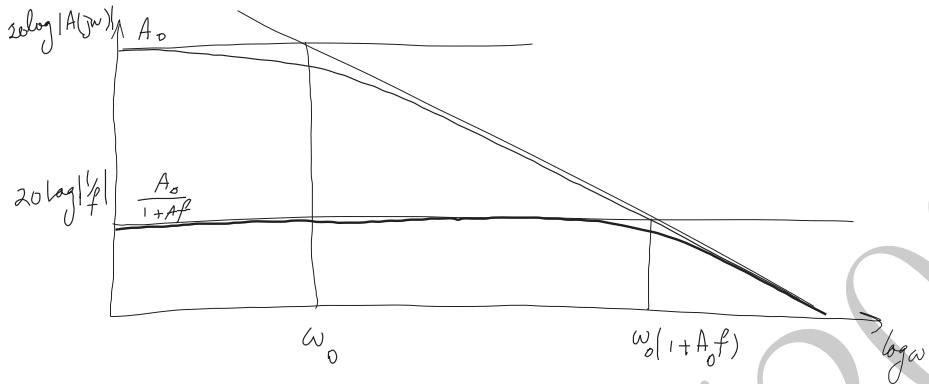


Figure 5.11: Amplifier with an LTI system in the feedback path.

where the last expression is written as the product of the low-frequency response of the feedback system,  $A_0/(1 + A_0 f)$ , and the frequency dependent part of the response. The closed-loop transfer function is still a low pass, but with a new time constant,  $\tau' = \tau/(1 + A_0 f)$ , which is a factor of  $1 + A_0 f$  smaller than that of the original forward path, corresponding a bandwidth increase by the same factor, as illustrated in Figure 5.11.

As is evident from the calculated  $H(s)$  in the last example, there is no amplitude peaking in the transition region between the  $1/f$  and  $A(s)$  asymptotic responses for the first-order forward path. This is, however, not generally the case as we will see in Chapter 6. The next example shows this behavior in a transistor level case.

**Example 5.1.8 (BJT Shunt-Shunt Amplifier: Frequency Response)** Looking at the BJT shunt-shunt amplifier of Example 5.1.3 modified by adding an input resistance  $R_1$ , as shown in 5.12, where this time the input is a voltage source,  $v_{in}$ , in series with a resistor,  $R_1$ . Let us consider the effect of the base-collector capacitor,  $C_\mu$ , and the base-emitter capacitor,  $C_\pi$ , on the bandwidth with and without the feedback resistor,  $R_2$ . Let us assume that the transistor collector current and hence its transconductance remains the same in both cases.

For the open-loop case, ( $R_2 \rightarrow \infty$ ), we can easily determine the low-frequency gain as the product of the voltage divider ratio formed by  $R_1$  and  $r_\pi$  times the intrinsic common-emitter gain, i.e.,

$$A(0) \equiv \frac{v_{out}}{v_{in}} = -\frac{r_\pi}{r_\pi + R_1} \cdot g_m R_3 \quad (5.10)$$

The 3dB bandwidth,  $\omega_h$ , can be estimated using the zero-value time-constants (ZVTs) using (4.42) of Chapter 4. As our analysis of the common-emitter stage in Example 4.3.3 showed the two time constant associated with  $C_\mu$  and  $C_\pi$  are

$$\tau_\pi^0 = C_\pi(r_\pi \parallel R_1) \quad \tau_\mu^0 = C_\mu[(1 + g_m R_3)(r_\pi \parallel R_1) + R_3]$$

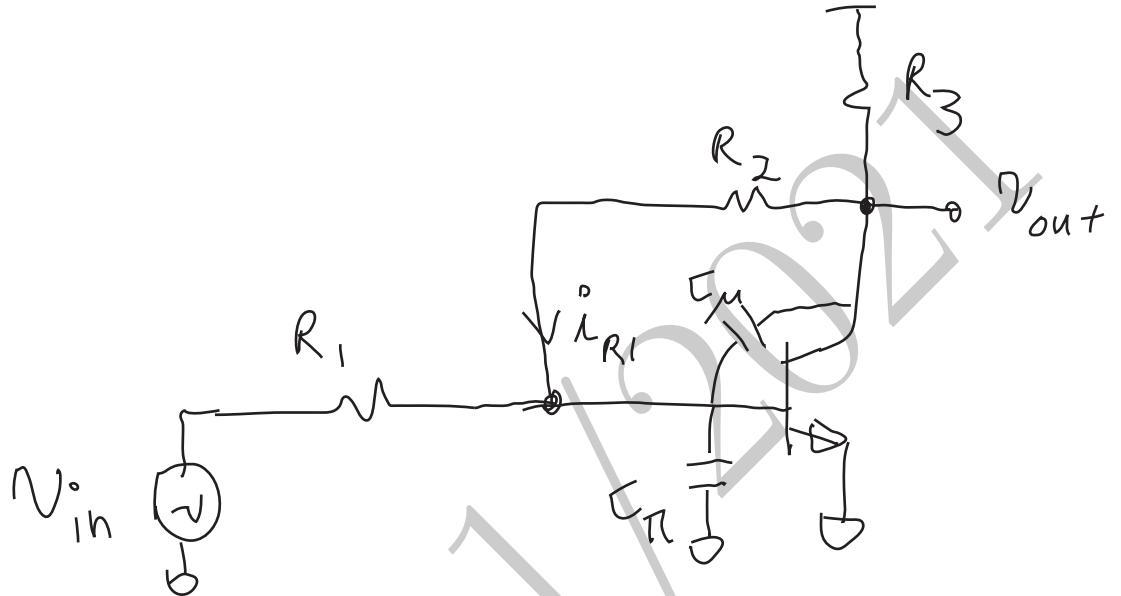


Figure 5.12: A BJT shunt-shunt feedback amplifier.

which leads to

$$\omega_h \approx \frac{1}{\tau_\pi^0 + \tau_\mu^0} \quad (5.11)$$

The reintroduction of the feedback resistor,  $R_2$ , lowers the low-frequency gain. Although we can easily calculate the new time-constants, for the sake of this example we just make the observation that  $R_2$  simply lowers the time constants  $\tau_\pi^0$  and  $\tau_\mu^0$ . At this stage it is easy to see why  $\tau_\mu^0$  is lowered. It is simply because the new resistor  $R_2$  is in parallel with the ZVT resistance seen by  $C_\mu$  before and thus lowers the total resistance it sees. The impedance seen by  $C_\pi$  is also reduced due to feedback. One intuitive way to see this is the Miller effect discussed in Subsection 4.2.3. We will see in more details later in Example 5.3.4 of Section 5.3 that the impedance seen by  $C_\pi$  is also reduced due to feedback, hence increasing the bandwidth.

We noticed in the previous example that for the first order system, the bandwidth increases by the same factor that the gain was reduced. Let us see if this observation about the first-order system can guide us in a numerical example.

Now, let us consider a numerical example. Let us assume transconductance of  $g_m = 40mS$ ,  $\beta$  of 100,  $C_\mu$  of  $100fF$ , and  $C_\pi$  of  $500fF$  for the bipolar transistor. For  $R_1 = 1k\Omega$ ,  $R_2 = 5k\Omega$ , and  $R_3 = 1k\Omega$ , the AC analysis in SPICE with and without  $R_2$  produces the small-signal transfer functions shown in Figure 5.13. The simulations predict an open-loop ( $R_2 \rightarrow \infty$ ) low frequency gain of 28.6 (29.1dB) and a 3dB bandwidth of 47MHz. matching the numerical

## ♦ Numerical Example ♦

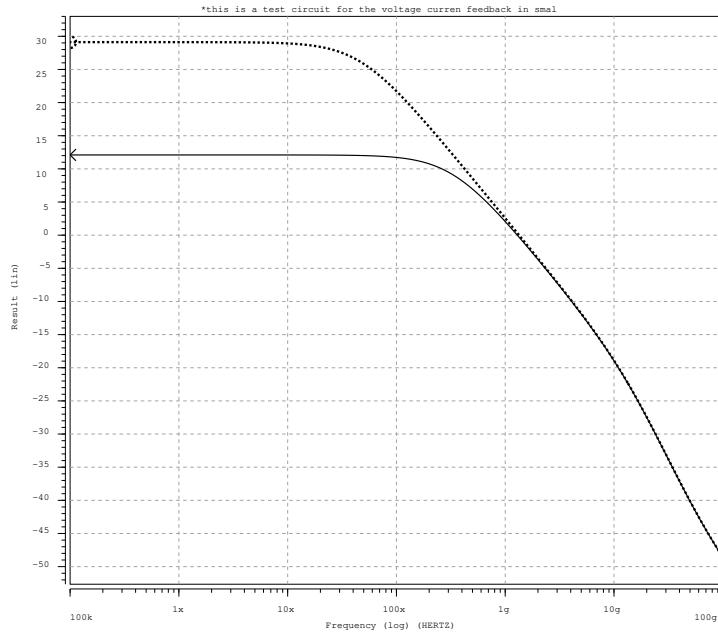


Figure 5.13: The ac transfer function of the shunt-shunt amplifier of 5.12 with and without the feedback resistor,  $R_2$ .

*results obtained from the analytical expressions of (5.10) and (5.11).*

*For the closed-loop system with  $R_2 = 5\text{k}\Omega$ , we have a low-frequency gain of 4.03 (12.1dB) with a 3dB bandwidth of 331MHz. The gain is reduced by a factor of  $28.6/4.03 \approx 7.09$ . At the same time the bandwidth is increased by a factor of  $331\text{MHz}/47\text{MHz} \approx 7.04$ . The feedback's reduction of the gain results in an almost equivalent increase in the bandwidth, as can be seen in Figure 5.13.*

### 5.1.2 Feedback Network with Linear Dynamics

The feedback network can have dynamics in general. In this subsection we will assume a linear time-invariant (LTI) network in the feedback path with a frequency domain transfer function,  $F(s)$ , as shown in Figure 5.14.

Maintaining the infinite forward gain assumption, the transfer functions of

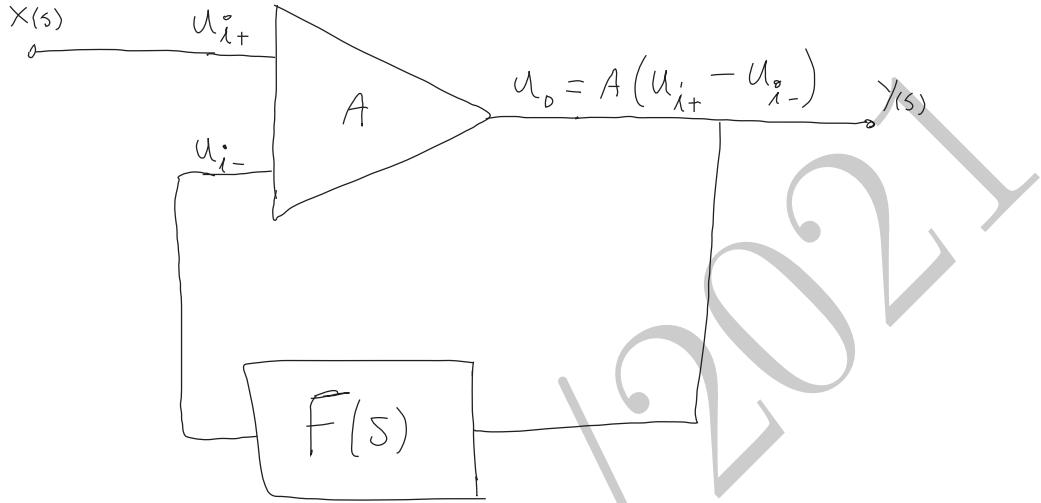


Figure 5.14: Amplifier with an LTI system in the feedback path.

(5.2) and (5.3) are still valid. Therefore, for an ideal forward path, we have<sup>8</sup>

$$H_\infty(s) = \frac{1}{F(s)} \quad (5.12)$$

which is the *inverse* transfer function of the feedback path in frequency domain. In other words, if the closed-loop system of Figure 5.14 is followed by a system with a transfer function  $F(s)$ , the overall system will have a unity transfer function for very large forward gain.

This is a useful feature and can be used to create transfer functions whose frequency-domain inverse are easier to implement, for instance, making a differentiator by using an integrator in feedback or creating a single zero transfer function with a first order pole as the feedback network.

**Example 5.1.9 (Inverting Op-Amp Low Pass Filter)** Considering the operation amplifier in inverting configuration with an additional capacitor in the feedback path, as illustrated in Figure 5.15. Now we notice that although the applied quantity at the input is a voltage ( $v_{in}$ ), the quantities that actually get subtracted ( $u_{i+}$  and  $u_{i-}$ ) are in fact the currents at the input node of the operational amplifier. So for our feedback viewpoint the input is really  $i_{i+}$ . However,

<sup>8</sup>In general, for a finite forward gain, the feedback transfer function can be written as the ratio of two polynomials in  $s$ , ( $G(s) = N(s)/D(s)$ ), the transfer function can be expressed as,

$$H(s) = \frac{A}{1 + AF(s)} = \frac{AD(s)}{D(s) + AN(s)}$$

It is noteworthy that the zeros of the closed-loop transfer function are the *poles* of the feedback network. In addition to this, in the case of  $A \rightarrow \infty$ , the poles of the closed-loop transfer function are also the same as the *zeros* of the feedback network as expected from (5.12).

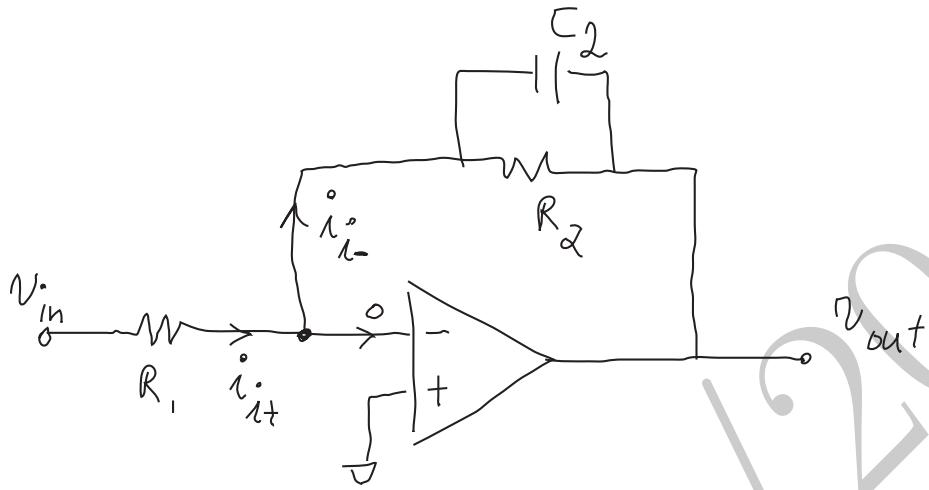


Figure 5.15: Op-Amp in inverting gain configurations with an RC network in the feedback path.

for large amplifier gain it is simply equal to  $v_{in}/R_1$  since the negative amplifier input is at ground (by asymptotic equality of the two op-amp inputs). The returned current subtracted from  $i_{i+}$  is the current through  $Z_2$  (the parallel combination of  $R_2$  and  $C_2$ ), namely,  $i_{i-}$  in Figure 5.15. According to the asymptotic equality principle for large op-amp gain, these currents are equal, hence we can simply write

$$i_{i+} = \frac{v_{in}}{R_1} = -\frac{v_{out}}{Z_2} = -\frac{v_{out}}{R_2}(1 + R_2Cs) = i_{i-}$$

which results in

$$\frac{v_{out}}{v_{in}} = -\frac{R_2}{R_1} \cdot \frac{1}{1 + R_2Cs}$$

which produces a low pass characteristic. While we have already arrived at the transfer function, it is instructive to calculate  $f$  and verify its relationship with  $H_\infty$ . The feedback block has  $v_{out}$  as its input and  $i_{i-}$  as its output, hence it is the admittance of the parallel combination of  $R_2$  and  $C$ , i.e.,

$$f \equiv \frac{i_{i-}}{v_{out}} = \frac{1 + R_2Cs}{R_2}$$

which is consistent with

$$H_\infty = \frac{v_{out}}{i_{i+}}|_{k \rightarrow \infty} = 1/f$$

Note that the feedback network is similar to that of Figure 4.5a which is a high pass network.

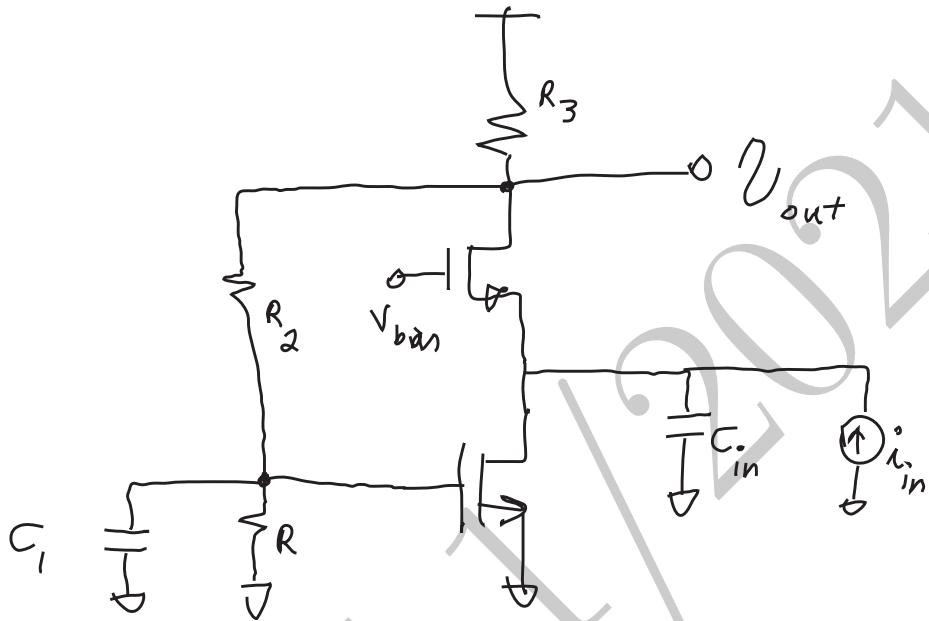


Figure 5.16: The MOS shunt-shunt feedback circuit with two capacitors.

Here is a transistor based example:

**Example 5.1.10 (MOS Shunt-Shunt Amplifier with low-pass feedback path)**  
*In the shunt-shunt stage of Example 5.1.5 shown in Figure 5.7, let us assume that there is a capacitance  $C_1$  between the gate of  $M_1$  and ground in parallel with  $R_1$ , which could be an explicit capacitor added by design or the gate-source capacitance of  $M_1$ . This capacitor turns the feedback network into a low-pass filter. Also assume that there is a second capacitor,  $C_{in}$  at the input in parallel with  $i_{in}$  as shown in Figure 5.16. Referring to the parallel combination of  $R_1$  and  $C_1$  as  $Z_1(s)$ , which is*

$$Z_1(s) = R_1 \parallel \frac{1}{C_1 s} = \frac{R_1}{1 + R_1 C_1 s},$$

*we can calculate the frequency-dependent  $H_\infty(s)$ . Again, for  $k \rightarrow \infty$ , we have  $i_{s2} = 0$  and  $v_{gs2} = 0$  and therefore the small-signal input voltage,  $v_{in} = 0$ . In this case, there will be no current flowing through the input capacitor,  $C_{in}$ , either. So we still have,  $i_{d1} = i_{in}$ . Replacing  $R_1$  with  $Z_1(s)$  in the calculations*

of Example 5.1.5, we obtain,

$$\begin{aligned} H_\infty(s) \equiv \frac{v_{out}}{i_{in}}|_{k \rightarrow \infty} &= \frac{1}{g_{m1}} \left[ 1 + \frac{R_2}{Z_1(s)} \right] \\ &= \frac{1}{g_{m1}} \left( 1 + \frac{R_2}{R_1} \right) \cdot [1 + (R_1 \parallel R_2)C_1 s] \\ &= H_\infty(0) \cdot (1 + \tau_z s) \end{aligned} \quad (5.13)$$

where  $\tau_z = (R_1 \parallel R_2)C_1$  and  $H_\infty(0)$  is the low-frequency asymptotic transfer function defined in (5.7).

It is interesting that the asymptotic transfer function can be characterized by a single zero which is the frequency domain inverse of a single-pole low-pass transfer function of the feedback network. The trans-impedance of the stage resembles that of an inductor and a resistor in series because of the capacitor in the feedback loop. Also note that the asymptotic transfer function is independent of  $C_{in}$ , as the asymptotic equality principle forces the input voltage to zero (or equivalently the input resistance is zero in the limit, therefore the time constant associated with  $C_{in}$  is zero.)

### 5.1.3 Forward Path Nonlinearity Reduction

Online YouTube lecture:

#### Effect of Feedback on Nonlinearity

So far we implicitly assumed that the forward path is linear. Interestingly, this is not necessary to obtain a relatively linear closed-loop input-output transfer function, as long as the amplifier gain is a high gain. To demonstrate this let us assume that the a memoryless nonlinearity,  $g(\cdot)$ , is present in the forward path, while we have a scalar linear feedback network,  $f$ , as depicted in Figure 5.17. In this case, the output,  $y$ , and the input,  $x$ , are related through a nonlinear relation,

$$y = g[A(x - u_{i-})] = g[A(x - fy)]$$

which cannot be analytically solved for  $y$ . However, we notice that we can solve for  $x$  in terms of  $y$ , i.e.,

$$x = fy + \frac{g^{-1}(y)}{A} \quad (5.14)$$

For a large forward gain ( $A \rightarrow \infty$ ), the second term diminishes and the transfer function simply reduces to that of (5.3) in the limit. Hence, as long as the gain in the forward path is large and the feedback network is linear, a nonlinearity in the forward path does not appear in the final transfer characteristic resulting in a linear closed-loop behavior<sup>9</sup>. This is an important result as it suggests a

<sup>9</sup>The implicit assumption here is that the nonlinear input-output characteristic,  $g(\cdot)$  is not zero for non-zero input values, as in such a case,  $g^{-1}(y)$  would be infinity and we need to resolve an indeterminant case.

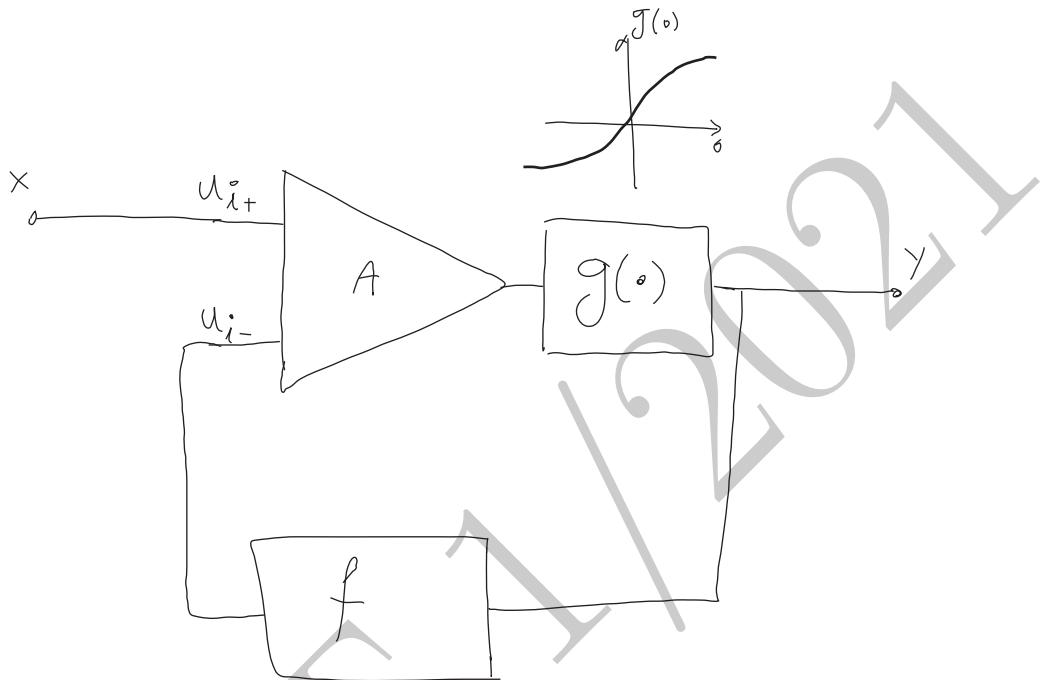


Figure 5.17: The effect of the amplifier nonlinearity in the presence of feedback.

practical way to overcome the inevitable nonlinearity of gain blocks and was in fact the primary reason Harold Black invented *negative feedback* as disclosed in his 1937 patent<sup>10</sup>.

**Example 5.1.11 (Cubic nonlinearity)** A cubic root nonlinearity,  $g(x) = \sqrt[3]{x}$ , in the forward path of the Figure 5.17, will have a limiting effect on larger signal amplitudes and introduces distortion in the signal. In this example, we will see how the large signal  $V_{in}-V_{out}$  transfer characteristic becomes more linear as the forward gain is increased. In this case, (5.14) reduces to:

$$x = fy + \frac{y^3}{A} \quad (5.15)$$

while this cubic equation can be solved analytically for  $y$  in terms of  $x$ , it is more instructive to look at it numerically. Let us assume that the feedback factor is  $f = 0.2$ , we will consider the system with the forward gains of  $A = 10^2$ ,  $A = 10^4$ , and  $A = 10^6$ . The open-loop ( $f = 0$ ) and closed-loop transfer functions for these three values of  $A$  are shown in Figure 5.18. It can be easily seen that increasing the gain, results in a more linear response. In the limit of  $A \rightarrow \infty$ , the response will be completely linear.

---

<sup>10</sup>U.S. Patent 2,102,671

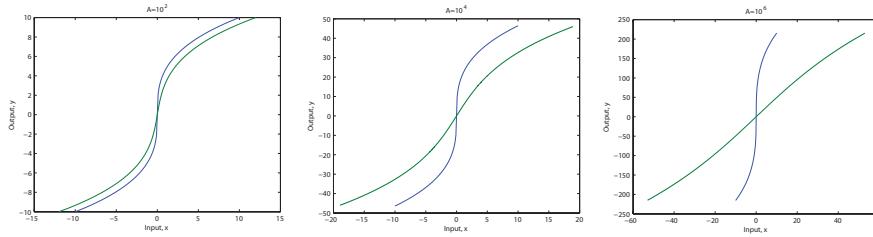


Figure 5.18: The effect of the amplifier gain on nonlinearity suppression in the presence of feedback shown for feedback factor of  $f = 0$  and  $f = 0.2$  for a)  $A = 10^2$ , b)  $A = 10^4$ , and c)  $A = 10^6$ .

Here is transistor level example:

**Example 5.1.12 (BJT shunt-shunt Amplifier: Large Signal Behavior)**  
*In this example, we will determine the low-frequency large-signal  $V_{in}$ - $V_{out}$  characteristic of the modified BJT shunt-shunt amplifier of Example 5.1.8 with an input resistance  $R_1$ , as shown in Figure 5.13. Since we are only interested in the low-frequency large-signal behavior of this amplifier, we will ignore the capacitors,  $C_\pi$  and  $C_\mu$  here.*

Assuming that the base current is negligible compared to the currents in  $R_1$  and  $R_2$  (i.e.,  $\beta \gg 1$ ), the currents through  $R_1$  and  $R_2$  are approximately equal<sup>11</sup>, i.e.,

$$\frac{V_{in} - V_{be}}{R_1} \approx \frac{V_{be} - V_{out}}{R_2} \quad (5.17)$$

that we can solve for  $V_{be}$  approximately,

$$V_{be} = V_{BE} + v_{be} \approx \frac{R_1 V_{out} + R_2 V_{in}}{R_1 + R_2} \quad (5.18)$$

where  $V_{BE}$  is the quiescent base-emitter voltage,  $v_{be}$  is the deviation from the operation point, and hence  $V_{be}$  is the overall base-emitter voltage.

The deviation from the operation point is amplified by the very steep exponential behavior of the transistor, therefore for the output to remain between the supply and ground, it has to be small. In other words, under stable operation,

<sup>11</sup>The circuit can be analyzed exactly by applying the KCL at the base and the collector of the transistor in conjunction with  $I_c = \beta I_b$  from (1.43) and the exponential behavior of the transistor, i.e.,  $I_c = I_s \exp(V_{be}/V_T)$  which was given in (1.38). This way we obtain:

$$\begin{aligned} I_c &= I_s \exp\left(\frac{V_{be}}{V_T}\right) = \frac{V_{be} - V_{out}}{R_2} + \frac{V_{CC} - V_{out}}{R_3} \\ I_b &= \frac{I_s}{\beta} \exp\left(\frac{V_{be}}{V_T}\right) = \frac{V_{in} - V_{be}}{R_1} - \frac{V_{be} - V_{out}}{R_2} \end{aligned} \quad (5.16)$$

which is challenging to solve for  $V_{out}$  in terms of  $V_{in}$  analytically.

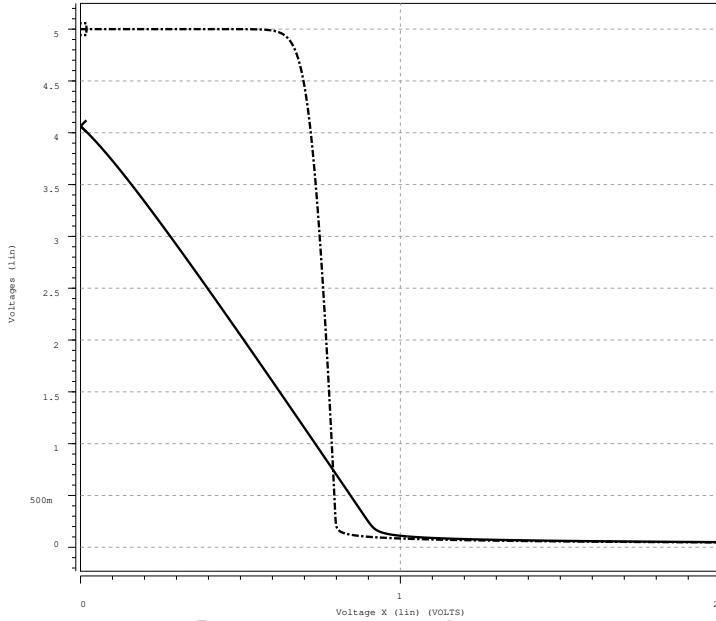


Figure 5.19: The *large-signal*  $V_{in}$ - $V_{out}$  of the shunt-shunt amplifier of Figure 5.12 with and without the feedback resistor,  $R_2$ .

$v_{be} \rightarrow 0$ , in (5.18) which then can be solved for  $V_{out}$  to provide

$$V_{out} = -\frac{R_2}{R_1}V_{in} + \left(1 + \frac{R_2}{R_1}\right)V_{BE,on} \quad (5.19)$$

which shows a linear dependence (with a constant dc offset) between the large-signal  $V_{in}$  and  $V_{out}$  unlike the case without the feedback resistor,  $R_2$ , where there is an exponential nonlinear dependence between the output and the input.

♦ Numerical Example ♦

Now, let us consider a numerical example. Let us assume a supply voltage of  $V_{CC} = 5V$  and a bipolar transistor with  $I_S = 10^{-15}A$  and  $\beta = 100$ . For  $R_1 = 1k\Omega$ ,  $R_2 = 5k\Omega$ , and  $R_3 = 1k\Omega$ , based on (5.19) we expect a linear large-signal  $V_{in}$ - $V_{out}$  characteristic with a slope of  $R_2/R_1 = 5$  crossing the y-axis at approximately<sup>12</sup>,  $(1 + 5) \times 0.7V = 4.2V$ . DC SPICE analysis of the circuit with and without  $R_2$  produces the large-signal  $V_{in}$ - $V_{out}$  characteristics shown in Figure 5.19. As can be seen the feedback results in a substantial improvement in the linearity of the amplifier response.

The linearization effect of feedback is useful when device nonlinearity produces undesirable distortion to the signal.

<sup>12</sup>The quiescent  $V_{BE}$  can be calculated directly to be  $0.675V$ .

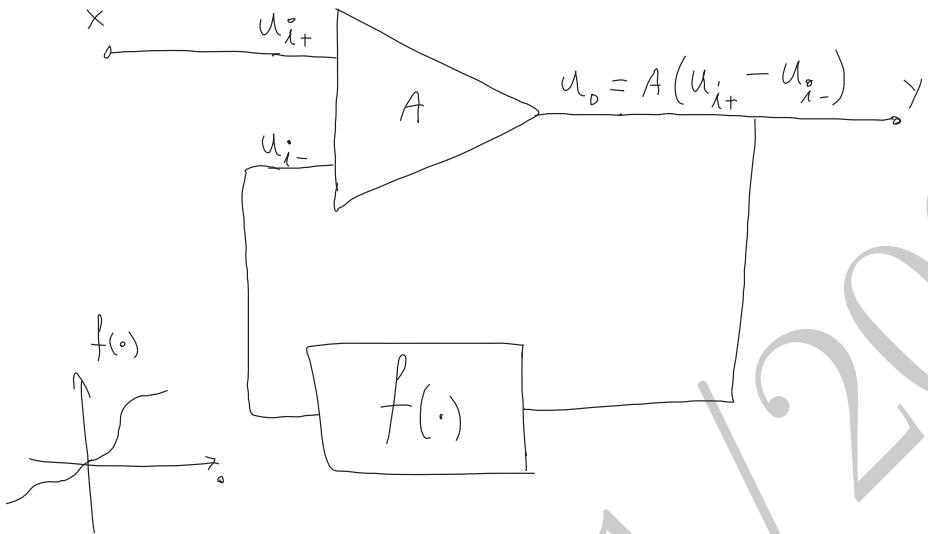


Figure 5.20: Amplifier with a memoryless nonlinearity in the feedback path.

### 5.1.4 Nonlinearity in the Feedback Path

Up until now, we have focused on linear networks in the feedback path. Now consider the case when the feedback network is formed by a memoryless nonlinearity represented by the function,  $f(\cdot)$ , as shown in Figure 5.20. In this case, we can calculate the relationship between the input and the output by noticing that  $u_{i-} = f(y)$ , and hence

$$y = A(x - u_{i-}) = A[x - f(y)]$$

which allows us to solve for  $x$  in terms of  $y$ ,

$$x = f(y) + \frac{y}{A} \quad (5.20)$$

that for  $A \rightarrow \infty$  reduces to  $x = f(y)$ , immediately implying

$$y = f^{-1}(x) \quad (5.21)$$

Interestingly, the asymptotic ( $A \rightarrow \infty$ ) nonlinear input-output transfer characteristic of the closed-loop system is the inverse function of the nonlinear feedback network's characteristic,  $f(\cdot)$ .

This behavior can be intuitively understood by noticing that in a negative feedback configuration, the feedback loop forces the output of the feedback network to follow the input closely (how close is determined by how large the loop gain is.) In the limiting case of infinite gain, the output (the input to the feedback network) must assume a value such that the output of the feedback network is the same as the input and hence  $y = f(x)$ .

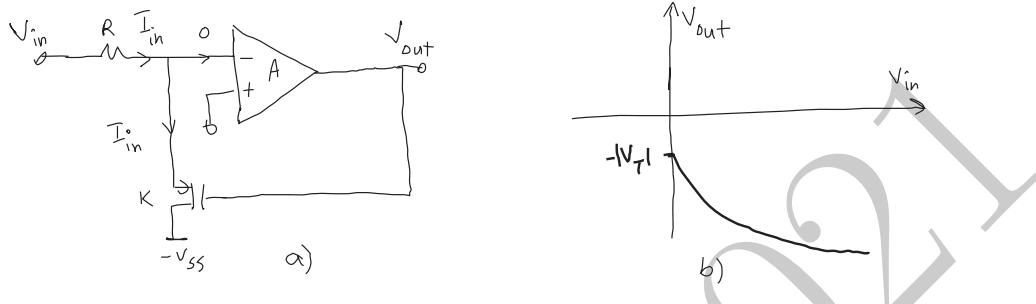


Figure 5.21: a) An ideal op-amp with a source follower in the feedback path, inverting its nonlinearity, b) its large-signal  $V_{in} - V_{out}$  transfer characteristic.

**Example 5.1.13 (Source Follower Feedback Network)** Consider the PMOS sources follower used as the negative feedback network in Figure 5.21a. Assuming that the op-amp has an infinite voltage gain, we use the asymptotic equality principle to conclude that the negative input of the op-amp is at ground potential, therefore the current through  $R_1$  directly flows into the source of  $M_1$ , i.e.,

$$I_{in} = \frac{V_{in}}{R_1} = I_D = k(-V_{out} + |V_T|)^2$$

which can be solved for  $V_{out}$  to provide

$$V_{out} = -\sqrt{\frac{V_{in}}{kR_1} - |V_T|}$$

This is the inverse of the quadratic transfer characteristic of the MOSFET itself, as shown in Figure 5.21b.

### 5.1.5 Sensitivity to Variations

Let us revisit the transfer function of the basic feedback system of Figure 5.1 given by (5.2). We saw earlier in (5.3) that if the forward gain is made very large this transfer function becomes independent of the forward loop gain and its behavior. This asymptotic behavior shows the general tendency of the feedback in the idealized gain. Now we show that even for a finite gain the general trend of desensitization to the forward gain behavior still holds. The easiest way to do so is to follow the same line as the original paper on negative feedback by Black [XXX]. First, we calculate the sensitivity of the closed-loop transfer function,  $H$ , given by (5.2) to the forward gain,  $A$ . This can be done by differentiating (5.2) with respect to  $A$ ,

$$\frac{dH}{dA} = \frac{1}{(1 + Af)^2} = \frac{1}{1 + Af} \cdot \frac{1}{A} \cdot \frac{A}{1 + Af} = \frac{1}{1 + Af} \cdot \frac{1}{A} \cdot H \quad (5.22)$$

which can be reordered as

$$\frac{dH}{H} = \frac{1}{1 + Af} \cdot \frac{dA}{A} \quad (5.23)$$

This clearly shows that the closed-loop transfer function,  $H$ , is desensitized to the open-loop forward gain,  $A$ , by a factor of  $1 + Af$ , which can be quite large. We also see that for  $A \rightarrow \infty$ , it has *no* sensitivity to gain variations, as suggested by (5.3). This is another manifestation of the fact that feedback reduces the dependence of the closed-loop system on the forward path's behavior.

Now let us determine the sensitivity of  $H$  to the feedback factor,  $f$ . Again this can be done by differentiating  $H$ , this time with respect to  $f$ ,

$$\frac{dH}{df} = -\frac{A^2}{(1 + Af)^2} = -\frac{Af}{1 + Af} \cdot \frac{1}{f} \cdot \frac{A}{1 + Af} = -\frac{Af}{1 + Af} \cdot \frac{1}{f} \cdot H \quad (5.24)$$

that directly results in

$$\frac{dH}{H} = -\frac{Af}{1 + Af} \cdot \frac{df}{f} \quad (5.25)$$

It is easy to see that there is very little desensitization to  $f$  for large  $Af$ . asymptotically, for  $A \rightarrow \infty$ , the sensitivity of  $H$  has the exactly same magnitude and an opposite polarity to that of  $f$ , again consistent with (5.3)<sup>13</sup>. As we can see any variations in the feedback factor directly translate to gain fluctuations. This is why the feedback networks are often implemented using simple passive components in such a way that  $f$  is determined by the ratio of the values of two passive elements (e.g.,  $R_2/R_1$  or  $C_2/C_1$ ). The use of the ratio of two elements has the added advantage of eliminating some of the systematic variations of the component values (e.g., with temperature or global process variations), as long as both passive elements are implemented the same way (i.e., same kind of resistors kept close to each other for temperature tracking).

The discussions in this section was primarily focused on the asymptotic loop behavior when the forward gain is large. In the next section, we will focus our attention on the effect of the finite gain and other non-idealities of a practical feedback circuit.

## 5.2 Finite Gain

In the previous section we calculated the asymptotic value of the closed-loop transfer function,  $H$ , when the scaled forward path gain was made infinite (via  $k \rightarrow \infty$ ), which was denoted as  $H_\infty$ . This is the most important parameter in a feedback system as it represents its ideal transfer function for infinite forward gain. In this section, we see how the result deviates from the ideal value for finite gain and present a correction factor. This correction factor is always there since we never have infinite gain, but its significance is a function of the magnitude of the forward gain and the feedback factor. One way to determine this correction factor is in terms of the quantity known as the *return ratio*.

<sup>13</sup>The reason for the minus sign is that  $H_\infty = 1/f$  so an increase in  $f$  results in a *reduction* in  $H$  and vice versa.

### 5.2.1 Return Ratio

Online YouTube lecture:

[\*\*Return Ratio, Asymptotic Gain Formula, Direct forward transfer\*\*](#)

We can define a *return ratio* for each dependent (controlled) source in the circuit. The return ratio of a given controlled source is a measure of how much of the signal generated by that dependent source is returned to it due to the feedback. It quantifies the strength of the feedback from the point of view of the controlled source for which return ratio is measured. The return ratios calculated for different dependent sources in the circuit are not necessarily equal and generally speaking each dependent source will have a return ratio of its own. Therefore, the return ratio is *not* a global property or an invariant of the loop. We will define the *loop gain* in Section 5.4 which is not associated with any given dependent source and is in general different from the return ratio. Loop gain is indeed an invariant of the loop, independent of the point of measurement. We will discuss the difference between loop gain and return ratio in great detail in Section 5.4. For the rest of this section we will focus our attention on return ratio.

Because of the feedback, a dependent source is affected by its own signal. More accurately, the signal it is controlled by is affected by the signal it generates. To determine the return ratio we must first null all external independent sources, including the input,  $u_i$ . Then to calculate the return ratio for a given controlled source denoted by  $u_y$ , we must separate that dependent source from the rest of the circuit, and replace it with an *independent* one of the same kind (voltage or current), denoted by  $u_x$ . The negative of the ratio of the signal generated by the dependent source to that of the independent source is defined as the return ratio,

$$T \equiv -\frac{u_y}{u_x} \quad (5.26)$$

The minus sign is introduced to make the return ratio a positive quantity for a negative feedback loop

In practice, the above procedure can be most easily performed by introducing an additive independent source  $u_z$  to  $u_y$  to generate the  $u_x$ . This is shown in Figure 5.22b and d for the dependent current and voltage source of Figure 5.22a and c, respectively. The dependent source,  $u_y$ , is proportional to  $u_a$  which is simply another signal (voltage or current) somewhere in the circuit, with a constant gain,  $g$ , i.e.,

$$u_y = gu_a \quad (5.27)$$

Note that introduction of the independent source  $u_z$  in Figure 5.22b and d does introduce an extra degree of freedom that can be used to set  $u_x$  to any desired value, since,

$$u_x = u_y + u_z \quad (5.28)$$

As a result, the introduction of  $u_z$  is *equivalent* to using an independent source  $u_x$  in place of  $u_y$  and monitoring  $u_y$ . Therefore,  $u_z$  should *never* appear in

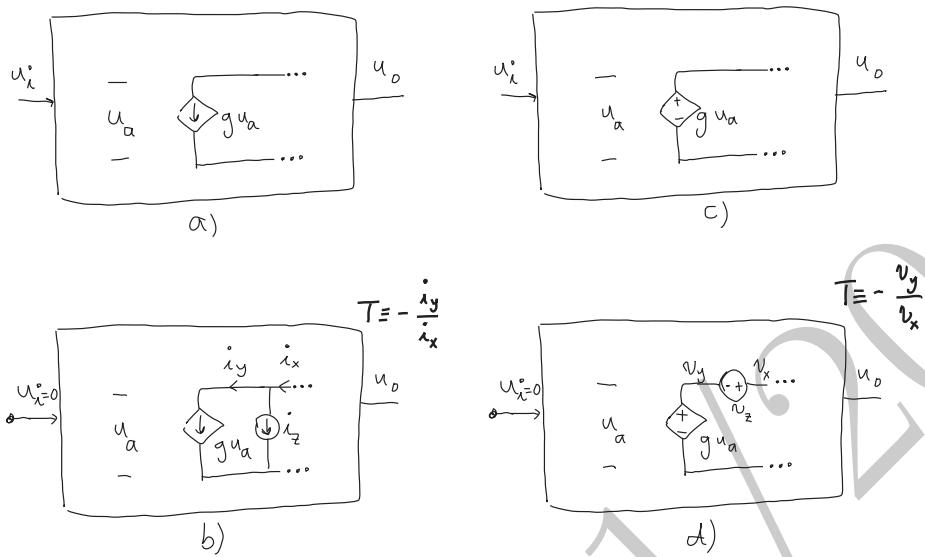


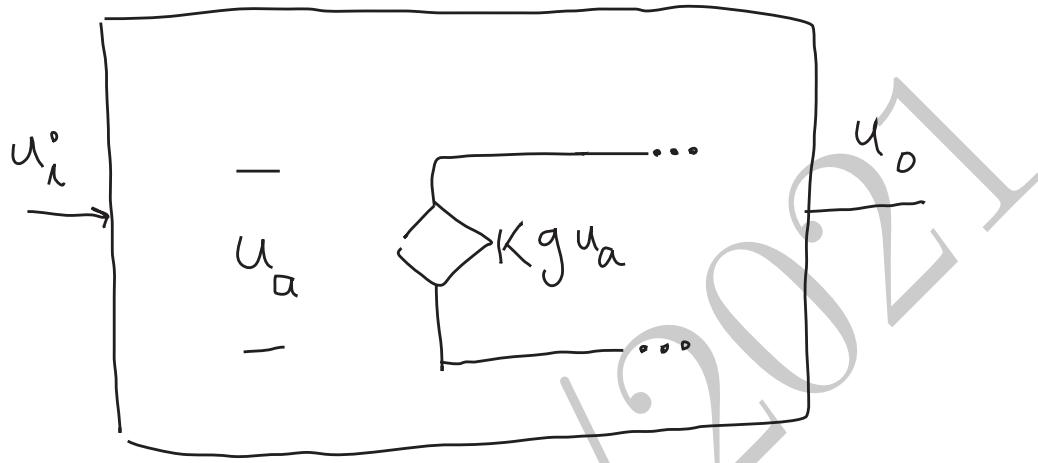
Figure 5.22: A general circuit with a) a dependent current source,  $gu_a$ , b) an independent current source,  $i_z$ , in parallel with  $gu_a$  for return ratio calculation, c) a dependent voltage source,  $gu_a$ , d) an independent voltage source,  $v_z$ , in series with  $gu_a$  for return ratio calculation.

the calculations directly (as will be seen in all the subsequent examples). It is just a convenient circuit means to create an independent  $u_x$  and monitoring the returned signal  $u_y$  without tearing the circuit apart. It merely makes it more convenient for circuit hand analysis and allows calculations to be done directly on the circuit diagrams in most cases. As can be seen from Figures 5.22b and d, a test current source,  $i_z$ , is placed in parallel with a dependent current source and a test voltage source,  $v_z$ , is applied in series with a dependent voltage source.

The return ratio can be determined for any dependent source in the circuit. However, there is no guarantee that different dependent sources will have the same return ratio and in general they do not. Therefore, it is only meaningful to define a return ratio for a given dependent source.

### 5.2.2 Asymptotic Transfer Function Formula

The above definition of return ratio allows us to express the transfer function in terms of its limiting (asymptotic) values. Consider a dependent source,  $u_y$ , somewhere inside the circuit in question. This dependent source could be either a voltage or current source, as depicted in Figure 5.22. Its value is proportional to another voltage or current in the circuit, generally shown as  $u_a$ , with a gain of  $g$ , as described by (5.27). To calculate the return ratio we introduce an *independent* source,  $u_z$ , that directly adds to the controlled source,  $u_y$  to

Figure 5.23: The reference dependent source scaled by a factor  $k$ .

produce  $u_x = u_y + u_z$  according to (5.28) and depicted in Figure 5.22. Note that if the controlled source ( $u_y$ ) is a current source,  $u_z$  will have to be a current source in parallel with it, and conversely if  $u_y$  represents a voltage source,  $u_z$  will be a voltage source in series with it, as shown in Figure 5.22b and d, respectively.

We can scale this source by a factor  $k$ , such that  $u_y = k u_a$ , as shown in Figure 5.23. This allows us to evaluate the response of the system in general under different conditions.  $k = 1$  corresponds to the nominal gain condition, where the circuit behaves as it did before scaling. By using other values of  $k$  we can evaluate the behavior of the circuit when the dependent source in question is very strong ( $k \rightarrow \infty$ ) or very weak ( $k = 0$ ).

Now in the most general sense we have two independent inputs to this network: the original input,  $u_i$ , and the test source,  $u_z$ , which together with  $u_y$  determine  $u_x$  through (5.28). For these two degrees of freedom we can choose any two linearly independent variables. Since we are more interested in  $u_x$ , we choose  $u_i$  and  $u_x$  to be our independent variable. This allows us to express the actual output,  $u_o$ , and the control variable for the dependent source,  $u_a$ , in terms of  $u_i$  and  $u_x$  as,

$$u_o = Au_i + Bu_x \quad (5.29a)$$

$$u_a = Cu_i + Du_x \quad (5.29b)$$

First, let us solve (5.29) in the absence of the test source, i.e., for  $u_z = 0$  to obtain the closed-loop transfer function,  $H$ . In this case, according to (5.28), we have  $u_x = u_y$  and can write

$$u_o = Au_i + Bu_y \quad (5.30a)$$

$$u_a = Cu_i + Du_y \quad (5.30b)$$

which together with (5.27) can be solved to obtain the transfer function in the absence of a test source<sup>14</sup>,

$$H \equiv \frac{u_o}{u_i}|_{u_z=0} = A + \frac{kgBC}{1 - kgD} \quad (5.31)$$

which can be reordered as

$$\begin{aligned} H \equiv \frac{u_o}{u_i}|_{u_z=0} &= \left(A - \frac{BC}{D}\right) \cdot \frac{-kgD}{1 - kgD} + A \cdot \frac{1}{1 - kgD} \\ &= H_\infty \cdot \frac{-kgD}{1 - kgD} + H_0 \cdot \frac{1}{1 - kgD} \end{aligned} \quad (5.32)$$

expressed in terms of  $H_\infty \equiv H|_{k \rightarrow \infty} = A - BC/D$  and  $H_0 \equiv H|_{k=0} = A$  which are the asymptotic values the transfer function assumes for  $k \rightarrow \infty$  and  $k = 0$ , respectively, in (5.31). Note that these are the values that the transfer functions assumes when the reference dependent source is made very strong or very weak.

The next step is to evaluate the return ratio in the circuit using (5.29). To determine the return ratio we must null the external independent source ( $u_i = 0$ ) in the presence of the test source  $u_z$  and evaluate the ratio of the returned signal,  $u_y$ , to  $u_x$ . By definition, this is done under nominal gain condition, hence for this calculation  $k = 1$ . For  $u_i = 0$ , equation (5.29b) reduces to  $u_a = Du_x$  which together with (5.27) provides  $u_y = gDu_x$ . Using the definition of the return ratio in (5.26), we determine it to be

$$T \equiv -\frac{u_y}{u_x} = -gD \quad (5.33)$$

which allows us to express (5.32) in terms of  $T$ .

### ▼ Result ▼

Thus, the closed-loop transfer function can be expressed as

$$H = H_\infty \frac{T}{1 + T} + H_0 \frac{1}{1 + T}$$

(5.34)

where  $T$  is the return ratio with respect to the dependent source  $g$ . Also  $H_\infty$  and  $H_0$  are called the *asymptotic transfer function* and *direct forward transmission* with respect to the same dependent source  $g$ . It is defined as the value of the transfer function when the same controlled source for which the return ratio was evaluated goes to infinity, ( $k \rightarrow \infty$ ),

$$H_\infty \equiv H|_{k \rightarrow \infty} \quad (5.35)$$

which serves as a more general definition of the asymptotic transfer function discussed in Section (5.1). The most general definition of return ratio based on the determinants of the circuit will be given in subsection 5.2.4.

In general, even when the dependent source in question is zero ( $k = 0$ ), there could be some direct transmission from the input ( $u_i$ ) to the output ( $u_o$ )

---

<sup>14</sup>This is done by replacing  $u_y$  by  $kg u_a$  in (5.30) and solving (5.30b) for  $u_a$  in terms of  $u_i$  and plugging it back into (5.30a) to solve for  $u_o/u_i$ .

through other parasitic and/or intentional signal paths. One common way for this to happen is signal transmission in reverse via the feedback network itself, which is often made out of simple passive components for linearity and stability reasons. So even for zero forward gain (which corresponds to zero return ratio for a dependent source in the forward path) there will be some residual signal transmission through the feedback path. Although this transmission is usually smaller than that of the main path, it can be significant at higher frequencies where the main path gain drops and this direct transmission path could take over. The direct forward transmission is not limited to the feedback network and can also happen through the parasitics of the main path.

The parameter  $H_0$  captures this direct transmission, as it is the value the transfer function assumes when the reference dependent source is rendered ineffective by setting  $k = 0$ . It represents the direct transmission through the circuit without that dependent source. Hence we call it the *direct forward transmission* term defined as:

$$H_0 \equiv H|_{k=0} \quad (5.36)$$

It is easy to see that (5.34) reduces to  $H_\infty$  for a large return ratio ( $T \rightarrow \infty$ ) and to  $H_0$  when the return ratio is zero ( $T = 0$ ). Equation (5.34) can also be written as  $H_\infty$  times two multiplicative correction factors,

$$H = H_\infty \cdot D_T \cdot D_0 \quad (5.37)$$

where  $D_T$  is called the *loop correction factor* which accounts for the deviation from the asymptotic (infinite-gain) transfer function,  $H_\infty$ , due to the finite return ratio in the circuit and is defined as

$$D_T \equiv \frac{T}{1+T} = \frac{1}{1+\frac{1}{T}} \quad (5.38)$$

and  $D_0$  is referred to as the *direct forward transmission correction factor* and can be calculated from (5.34) and (5.38) to be

$$D_0 = 1 + \frac{H_0}{TH_\infty} \quad (5.39)$$

It is easy to see from (5.38) and (5.39) that both  $D_T$  and  $D_0$  approach unity for a large return ratio ( $T \rightarrow \infty$ ).

Note that  $D_T < 1$  for positive  $T$ . Also note that  $D_0$  can be greater than unity if  $H_0$  and  $H_\infty$  have same polarities.

It should be noted that in general using a different dependent source as a reference for these calculations results in different set of values for  $T$ ,  $H_\infty$ , and  $H_0$ . Nonetheless, (5.34) remains valid as long as *all three* quantities are measured with respect to the *same* dependent source.

It is obvious that if the direct forward transmission is not significant ( $H_0 \approx 0$ ), the second term in (5.34) can be ignored (i.e.,  $D_0 = 1$ ) and it reduces to

$$H \approx H_\infty \cdot \frac{T}{1+T} = H_\infty \cdot D_T \quad (5.40)$$

Several observation as in order regarding (5.34):

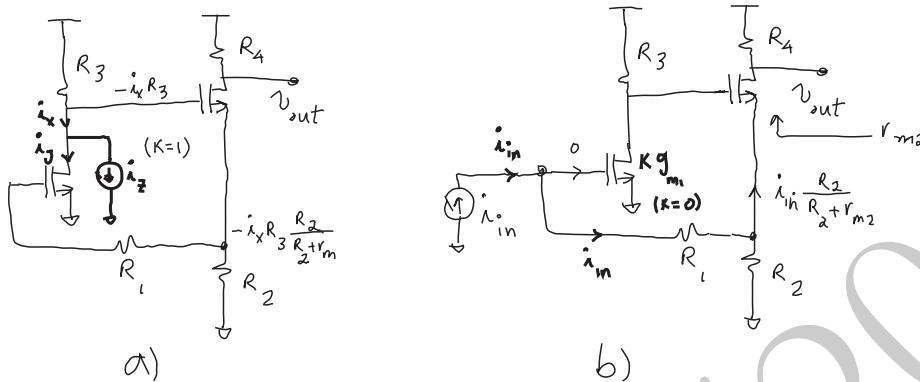


Figure 5.24: Calculation of a) the return ratio,  $T$ , and b) direct forward transmission,  $H_0$ , for the shunt-series feedback stage of Figure 5.6.

1. All three need to be evaluated for the same dependent source
2. The transfer function can be a gain or an impedance (similar to before)
3. This is a useful decomposition as it breaks the calculation of the transfer function into three simpler calculations, each with its own physical interpretation
4. Return ratio is not always equal to the loop gain, as we will see and discuss in Section 5.4

**Example 5.2.1 (MOS Shunt-Series Feedback Stage)** We calculated the asymptotic transfer function of the shunt-series feedback amplifier of Figure 5.6a with respect to the dependent current source of  $M_1$ , in Example 5.1.4. Now we will calculate the return ratio,  $T$ , and the direct forward transmission,  $H_0$ , with reference to the same dependent source. If we intend to use the asymptotic transfer function ( $H_\infty$ ) obtain in (5.6) in Example 5.1.4, we must calculate  $T$  and  $H_0$  with respect to the same dependent source that was used to obtain  $H_\infty$ , namely,  $g_{m1}$ .

In Figure 5.24a, we calculate the return ratio, by nulling the input ( $i_{in}$ ) and making  $i_x$  an independent source and calculating the returned current,  $i_y$ , with the aid of  $i_z$ . The current  $i_x$  is pulled out of  $R_3$  producing a voltage,  $-R_3 i_x$  at the gate of  $M_2$ . This voltage experiences the gain of the source follower formed by  $M_2$  and  $R_2$  which is determined by the voltage divider ratio between  $r_{m2}$  and  $R_2$  since no current flows through  $R_1$  at low frequencies. Therefore the voltage at the gate of  $M_1$  is the same as the source of  $M_2$ , which is

$$v_{g1} = -i_x R_3 \cdot \frac{R_2}{r_{m2} + R_2}$$

However, the returned current,  $i_y$ , is simply  $g_{m1}v_{g1}$ , hence,

$$T \equiv -\frac{i_y}{i_x} = g_{m1}R_3 \cdot \frac{R_2}{r_{m2} + R_2} = g_{m1}R_3 \cdot \frac{g_{m2}R_2}{1 + g_{m2}R_2} \quad (5.41)$$

which can be used to calculate the loop correction factor defined in (5.38),

$$D_T \equiv \frac{T}{1 + T} = \frac{g_{m1}g_{m2}R_2R_3}{1 + g_{m2}R_2(1 + g_{m1}R_3)} \quad (5.42)$$

which is slightly smaller than unity for typical values of transconductances and resistors.

To calculate the direct forward transmission with respect to the transconductance of  $M_1$ , we have to scale it by  $k$  and set  $k$  to zero, as shown in Figure 5.24b. Now, since the low-frequency input impedance of  $M_1$  is infinity, the input current  $i_{in}$  will flow through  $R_1$  and is then current divided between the source impedance of  $M_2$ , i.e.,  $r_{m2}$  and the resistor  $R_2$ , to produce the  $M_2$ 's source current  $i_{s2}$  which appears as a common gate<sup>15</sup> for this input and whose source and drain currents are equal at low frequencies,

$$i_{d2} = i_{s2} = -i_{in} \cdot \frac{R_2}{R_2 + r_{m2}}$$

However, this current flows through  $R_4$  to produce  $v_{out}$ , thus, the direct forward transmission is

$$H_0 \equiv \frac{v_{out}}{i_{in}}|_{k=0} = R_4 \cdot \frac{R_2}{R_2 + r_{m2}} = R_4 \cdot \frac{g_{m2}R_2}{1 + g_{m2}R_2} \quad (5.43)$$

that allows us to calculate  $D_0$  defined in (5.39),

$$D_0 \equiv 1 + \frac{H_0}{H_\infty T} = 1 + \frac{R_2}{g_{m1}R_3(R_1 + R_2)} \quad (5.44)$$

which is slightly greater than unity. This is because both  $H_\infty$  and  $H_0$  have the same signs, namely the parasitic signal transmission in reverse through the feedback network has the same polarity as the primary path whose transfer function is idealized by the asymptotic transfer function,  $H_\infty$ .

Now the complete transfer function can be calculated from (5.37) (or (5.34)) to be

$$\begin{aligned} H &= H_\infty \cdot D_T \cdot D_0 \\ &= R_4 \left( 1 + \frac{R_1}{R_2} \right) \cdot \frac{g_{m1}g_{m2}R_2R_3}{1 + g_{m2}R_2(1 + g_{m1}R_3)} \cdot \left[ 1 + \frac{R_2}{(R_1 + R_2)g_{m1}R_3} \right] \\ &= \frac{g_{m2}R_4 \cdot [R_2 + g_{m1}R_3(R_1 + R_2)]}{1 + g_{m2}R_2(1 + g_{m1}R_3)} \end{aligned} \quad (5.45)$$

which is approximately equal to  $H_\infty$ , for typical values.

---

<sup>15</sup>It is interesting to note that  $M_2$  appears as a (degenerate) common source for the main signal transmission, a source follower for the feedback signal, and a common gate for the direct forward transmission in this topology.

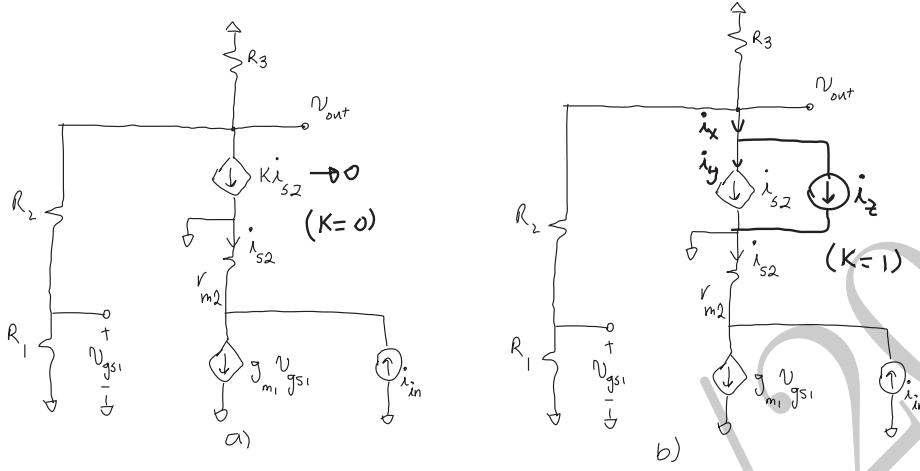


Figure 5.25: Calculation of a) the return ratio,  $T$ , and b) direct forward transmission,  $H_0$ , for the shunt-shunt feedback stage of Figure 5.7.

The final result for transfer function obtained in (5.45) could be obtained by direct application of nodal analysis. However, the above approach allows us to perform the analysis in a stepwise fashion being able to stop at the right time. Also the above approach is more conducive to design as once we know the desired  $H_\infty$  we can choose the resistors,  $R_1$ ,  $R_2$ , and  $R_4$  to give us the desired transimpedance, and view the impact of the finite gain via the loop correction factor,  $D_T$ , and that of the direct forward transmission through  $D_0$ . We would like both of these parameters to be as close to unity for the actual transfer function,  $H$ , to be close to the ideal (asymptotic) transfer function,  $H_\infty$ . One obvious way to achieve this is by increasing the return ratio (for example by increasing  $g_{m1}R_3$ ) which brings both  $D_T$  and  $D_0$  closer to unity.

**Example 5.2.2 (MOS Shunt-Shunt Amplifier: Direct Forward Transmission and Return Ratio)**  
*Now let us determine the direct forward transmission,  $H_0$ , and the return ratio,  $T$ , in the shunt-shunt stage of Figure 5.7. Since we have already calculated the asymptotic transfer function with respect to  $T$ -model dependent current source of  $M_2$  in (5.7) of Example 5.1.5, we must evaluate the  $H_0$  and  $T$  with respect to the same source to be able to reuse the  $H_\infty$  result of (5.7) in the asymptotic transfer function formula.*

*To calculate,  $H_0$ , we must set  $k$  to zero. In this case, there will be no signal transmission through  $M_2$  (assuming  $r_o \rightarrow \infty$ ), as can be seen from Figure 5.25a. Also there is no low-frequency reverse transmission from drain of  $M_1$  to its gate, hence the direct forward transmission is simply zero, i.e.,*

$$H_0 = \frac{v_{out}}{i_{in}}|_{k=0} = 0 \quad (5.46)$$

*which directly implies,  $D_0 = 1$ . Note that this would not be necessarily the case,*

for a different reference controlled source<sup>16</sup>. This also fails at higher frequencies due to the gate-drain capacitance ( $C_{gs1}$ ) of  $M_1$  which creates a reverse signal path at higher frequencies.

To calculate the return ratio with respect to the same dependent source, the input current,  $i_{in}$ , is nulled and an independent current source  $i_z$  in parallel with  $i_{s2}$  produces an independent current,  $i_x$ , as shown in Figure 5.25b. This current is divided between  $R_3$  and  $R_1 + R_2$ , so the current through  $R_1$  is simply  $-i_x R_3 / (R_1 + R_2 + R_3)$ . Thus the returned current,  $i_y$ , is equal to the drain current of  $M_1$  (since return ratio calculations are done for nominal gain conditions, i.e.,  $k = 1$ ) which is simply its gate voltage times,  $g_{m1}$ . Therefore, the return ratio is

$$T \equiv -\frac{i_y}{i_x} = g_{m1} R_3 \frac{R_1}{R_1 + R_2 + R_3} \quad (5.47)$$

Therefore, since  $H_0 = 0$  for this choice of the reference source, the complete low-frequency transfer function using (5.40) is simply,

$$H = H_\infty \cdot \frac{T}{1+T} = \frac{1}{g_{m1}} \left(1 + \frac{R_2}{R_1}\right) \cdot \frac{g_{m1} R_1 R_3}{R_1 + R_2 + R_3 + g_{m1} R_1 R_3} \quad (5.48)$$

which is close to  $H_\infty$  when the  $g_{m1} R_3$  product is large.

### Example 5.2.3 (MOS Shunt-Shunt Amplifier: High Frequency Behavior)

Now let us evaluate  $T$  and  $H_0$  for shunt-shunt feedback amplifier of Figure 5.16 in the presence of the input capacitor  $C_2$  and the gate capacitor,  $C_1$ . We have already determined the low-frequency asymptotic transfer function under these conditions in (5.13) of Example 5.1.10 to be

$$H_\infty(s) \equiv \frac{v_{out}}{i_{in}}|_{k \rightarrow \infty} = H_\infty(0) \cdot (1 + \tau_z s)$$

where  $\tau_z = (R_1 \parallel R_2) C_1$  is the zero time constant and  $H_\infty(0) = r_{m1}(1 + R_2/R_1)$  is the low-frequency asymptotic transfer function, as determined in Example 5.1.10.

The direct forward transmission,  $H_0$ , can be easily determined to be zero by noting that when  $k = 0$ , there is no path for the signal to flow from the input to the output<sup>17</sup>, as depicted in Figure 5.26a.

The return ratio with respect to the T-model dependent source of  $M_1$  can be determined by using the method of time-constants developed in Chapter 4, considering that for this calculation the output node is driven by an independent current source,  $i_x$ , as shown in Figure 5.26b. Therefore, the resistance seen by

<sup>16</sup>For instance, if we had chosen the  $\pi$ -model dependent current source of  $M_1$  ( $g_{m1} v_{gs1}$ ) as the reference, setting it to zero would have resulted in a non-zero,  $H_0$ , via the common-gate of  $M_2$ . Although calculating  $H_\infty$  and  $T$  with respect to  $g_{m1}$  is possible and produces a correct result in the asymptotic transfer function formulas, (5.37) or (5.34), it is not the best choice since it results in expressions for  $H_\infty$  and  $T$  that do not have as intuitive an interpretation, particularly at higher frequencies.

<sup>17</sup>This will not be true in the presence of the gate-source capacitance of  $M_1$ , namely,  $C_{gd1}$ , which provides a path for the input signal to get to the output through,  $R_2$ , when  $k = 0$ .

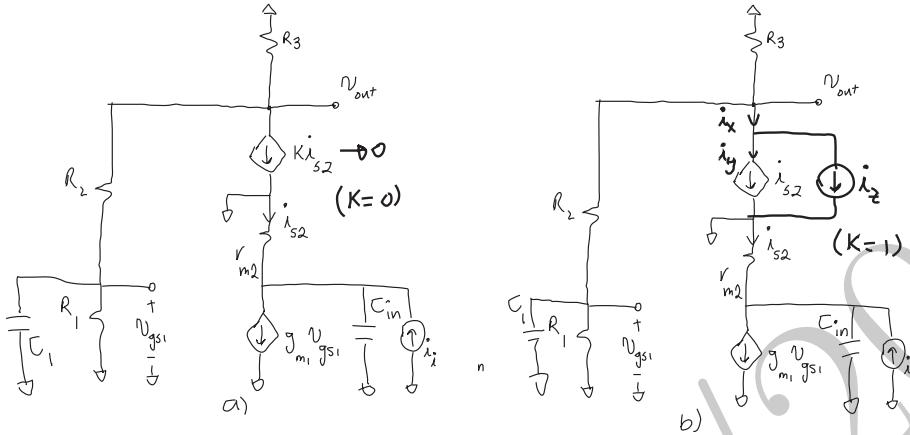


Figure 5.26: Calculation of a) the high frequency return ratio,  $T(s)$ , and b) direct forward transmission,  $H_0$ , for the shunt-shunt feedback stage of Figure 5.16.

$C_1$  when the independent source is nulled ( $i_x$  is open) is simply  $R_1 \parallel (R_2 + R_3)$ , hence,

$$\tau_1^0 = [R_1 \parallel (R_2 + R_3)]C_1 \quad (5.49)$$

The resistance seen by  $C_{in}$  is obviously the source resistance of  $M_2$ , namely,  $r_{m1}$ . Thus,

$$\tau_{in}^0 = r_{m2}C_2 \quad (5.50)$$

It is easy to see that in this case, the two time constants are uncoupled (i.e.,  $\tau_1^0 = \tau_2^0$ ) since shorting or opening of one of the capacitors does not change the resistance seen by the other one. Therefore each time constant simply represents a separate real pole time-constant. We also notice that the shorting of either  $C_1$  or  $C_2$  results in a zero return ratio, hence  $T^1$ ,  $T^2$ , and  $T^{12}$  are all zero and we do not have a zero in the transfer function. Noting that the low-frequency return ratio is already calculated in (5.47) in the previous example, we can express the return ratio as

$$\begin{aligned} T(s) &= \frac{g_{m1}R_1R_3}{R_1 + R_2 + R_3} \cdot \frac{1}{1 + [R_1 \parallel (R_2 + R_3)]C_1s} \cdot \frac{1}{1 + r_{m2}C_{in}s} \\ &= T(0) \cdot \frac{1}{1 + \tau_1 s} \cdot \frac{1}{1 + \tau_{in} s} \end{aligned} \quad (5.51)$$

where  $T(0)$  is the low-frequency return ratio given in (5.47), and  $\tau_1$  and  $\tau_2$  are short hand notations for the ZVT's since they are decoupled.

The transfer function can therefore be expressed as

$$H(s) = H(0) \cdot \frac{1 + \tau_z}{1 + b_1 s + b_2 s^2} \quad (5.52)$$

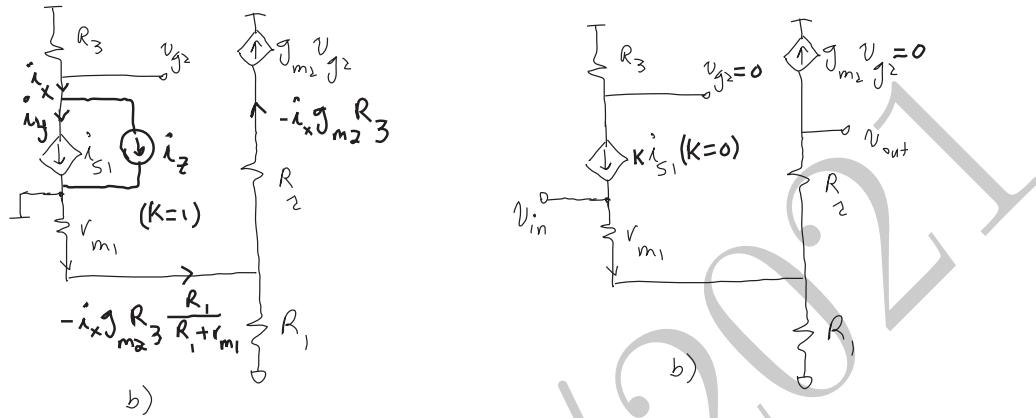


Figure 5.27: Calculation of a) the return ratio,  $T$ , and b) direct forward transmission,  $H_0$ , for the series-shunt feedback stage of Figure 5.4.

where  $H(0) = H_\infty(0)D_T(0) = H_\infty(0)T(0)/[1+T(0)]$  is the low frequency transfer function,  $\tau_z = (R_1 \parallel R_2)C_1$  is the zero time constant, and

$$b_1 = \frac{\tau_1 + \tau_2}{1 + T(0)}$$

$$b_2 = \frac{\tau_1 \tau_2}{1 + T(0)}$$

which shows that due to the feedback, the two time constants of the closed-loop system are not uncoupled and that for

$$T(0) > \frac{(\tau_1 - \tau_2)^2}{4\tau_1\tau_2} \quad (5.53)$$

we have a pair of complex conjugate poles. Also the LHP zero in the overall transfer function can result in peaking if it occurs before the two poles.

**Example 5.2.4 (MOS Series-Shunt feedback amplifier:Return Ratio)**  
 Going back to the MOS series-shunt amplifier of Example 5.1.2 shown in Figure 5.4a, this time we calculate the return ratio,  $T$ , and the direct forward transmission,  $H_0$ .

Since we determined the asymptotic transfer function,  $H_\infty$  in (5.4) with respect to the dependent current source of the T-Model for  $M_1$ , in Figure 5.4a, we should determine  $T$  and  $H_0$  with reference to the same dependent source to be able to use the asymptotic transfer function formula. To determine the return ratio, we apply a test current source,  $i_z$ , in parallel with it, i.e., between the drain and the gate of  $M_1$ , as shown in Figure 5.27a. The current  $i_x$  flows out of  $R_3$  generating a voltage  $-R_3 i_x$  at the gate of  $M_2$  and its drain current is  $-g_{m2} R_3 i_x$ . This current flows through  $R_2$  and the parallel combination of  $R_1$

and  $r_{m1}$ . Hence the source current of  $M_1$  is determined by the current divider formed by  $R_1$  and  $r_{m1}$ . Thus, the return ratio which is the negative of the of the returned current  $i_y = i_{s1}$  (since  $k = 1$ ) to  $i_x$  is

$$T \equiv -\frac{i_y}{i_x} = g_{m2}R_3 \cdot \frac{R_1}{R_1 + r_{m1}} = g_{m1}g_{m2}R_3(R_1 \parallel r_{m1}) \quad (5.54)$$

Now the direct forward transmission can be calculated by setting  $k$  to zero, as illustrated in Figure 5.27b. At low frequencies the only signal path is through  $M_1$  and the feedback network consisting of  $R_1$  and  $R_2$ . In this case,  $M_1$  and  $R_1$  form a source follower stage that is connected to the output via  $R_2$ . The small-signal gain between  $v_{in}$  and  $v_o$  in this case is determined by voltage divider between  $r_{m1}$  and  $R_1$ .

$$H_0 = \frac{v_o}{v_{in}}|_{k=0} = \frac{R_1}{R_1 + r_{m1}} \quad (5.55)$$

This calculation completes the parameters needed for determination of the complete transfer function using (5.37) (or (5.34)):

$$\begin{aligned} H &= H_\infty D_T D_0 \\ &= \left(1 + \frac{R_2}{R_1}\right) \cdot \frac{g_{m1}g_{m2}R_1R_3}{1 + g_{m1}R_1(1 + g_{m2}R_3)} \cdot \left[1 + \frac{R_1}{(R_1 + R_2)g_{m2}R_3}\right] \\ &= \frac{g_{m1}g_{m2}R_3(R_1 + R_2) + g_{m1}R_1}{1 + g_{m1}R_1(1 + g_{m2}R_3)} \end{aligned} \quad (5.56)$$

which is the same result that can be directly obtained from nodal analysis<sup>18</sup>.

Note that the nodal analysis produces the result in the form of the third line of the above equation, while the asymptotic transfer function analysis generates it in the more useful form of the second line. It is more useful because it separates different effects and allows the designer to start with  $H_\infty$ , which is primary part of the design and continue to improve the design further by taking into account the other two correction factors.

The return ratio can be calculate for any dependent source and as long as the exact same dependent source is used in determination of  $H_0$  and  $H_\infty$ , (5.34) can be used to determine the transfer function exactly. This concept can be seen in the following example, which is the same MOS series-shunt amplifier of Example 5.1.2 shown in Figure 5.4a that was analyzed in the last example. The only difference is that this time we use the  $\pi$ -model for  $M_1$  instead of the T-model which leads to different  $T'$ ,  $H'_0$ , and  $H'_\infty$  but the same final transfer function.

<sup>18</sup>To do so we can call the small-signal voltage of the source of  $M_1$   $v_1$  and the gate voltage of  $M_2$  is called  $v_2$ , we can write KCL at the source of  $M_1$ :

$$\begin{aligned} v_1 &= R_1(i_{s1} - i_{s2}) = R_1[g_{m1}(v_{in} - v_1) - g_{m2}v_2] \\ &= R_1[g_{m1}(v_{in} - v_1) + g_{m2}(v_{in} - v_1)] = R_1(g_{m1} + g_{m2})(v_{in} - v_1) \end{aligned}$$

we then note that  $v_o = v_1 - g_{m2}v_2 = v_o + g_{m1}g_{m2}R_L(v_{in} - v_1)$  which can be used to eliminate  $v_1$  in the previous equation to produce the transfer function of (5.56).

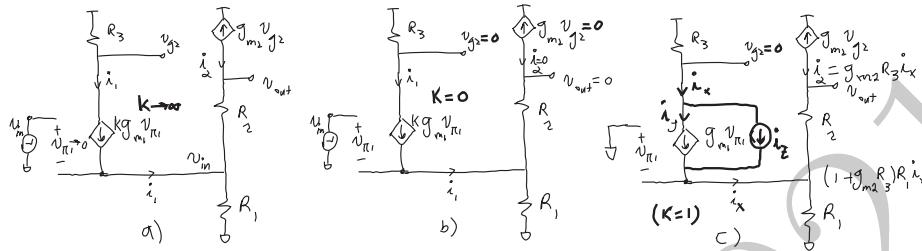


Figure 5.28: Calculation of a)  $H'_\infty$ , b)  $T'$ , and b)  $H'_0$ , for the series-shunt feedback stage of Figure 5.4 with respect to the dependent current source of the  $\pi$ -model for  $M_1$ .

Online YouTube lecture:

[Active feedback, dual feedback loop, Source dependence of return ratio examples.](#)

### Example 5.2.5 (MOS Series-Shunt feedback amplifier:Return Ratio (alternative source))

Using a  $\pi$ -model for the MOS series-shunt amplifier of Figure 5.4a, we have a small signal model shown in Figure 5.28a. This time we use the drain-source current source ( $g_{m1}v_{\pi1}$ ) as the reference source. First let us calculate  $H'_\infty$ . To have a finite output voltage when  $k' \rightarrow \infty$ , the gate-source voltage,  $v_{\pi1}$  must be zero, which indicates that the feedback forces the voltage across  $R_1$  to be  $v_{in}$  asymptotically. It also implies that the small-signal drain and source currents of  $M_1$  are equal. We will show this current as  $i_1$ . This indicates that the small signal gate voltage of  $M_2$  is  $-i_1R_3$  which in turn indicates that the current injected into  $R_2$  through the drain of  $M_2$ , is  $i_2 = i_1g_{m2}R_3$ . Since the total current through  $R_1$  is  $i_1 + i_2$ , we can write:

$$v_{in} = R_1(i_1 + i_2) = R_1(1 + g_{m2}R_3)i_1$$

Now, the output voltage  $v_o$  is the voltage drop across  $R_2$  plus the voltage drop across  $R_1$  that is simply  $v_{in}$ , thus,

$$v_o = v_{in} + R_2i_2 = v_{in} + i_1g_{m2}R_2R_3 = v_{in} \left[ 1 + \frac{g_{m2}R_2R_3}{R_1(1 + g_{m2}R_3)} \right]$$

which results in

$$H'_\infty \equiv \frac{v_o}{v_{in}}|_{k' \rightarrow \infty} = 1 + \frac{R_2}{R_1} \cdot \frac{g_{m2}R_3}{1 + g_{m2}R_3} \quad (5.57)$$

which is clearly different from the  $H_\infty$  calculated with respect to the dependent current source of the T-model for  $M_1$  obtained in (5.4).

Now let us calculate  $H'_0$ . To do so, we must set the dependent current source to zero, as shown in Figure 5.28b. It is easily seen that in this case, no signal leaks to the output and hence  $H'_0 = 0$ , again different from (5.55).

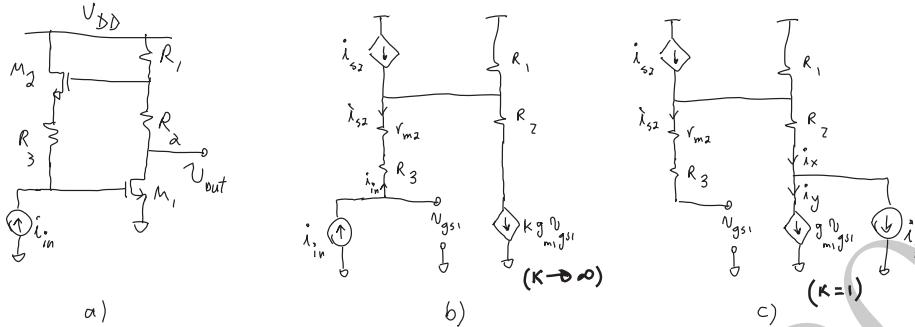


Figure 5.29: a) Cherry-Hooper stage, b) small-signal load and  $H_\infty$  calculation, c) calculation of the return ratio with respect to the dependent current source of the  $\pi$ -model for  $M_1$ .

The last step is to determine the return ratio with respect to this new dependent source. To do this we need to introduce a current source  $i'_z$  in parallel with it and measured the returned current  $i'_y$  as a function of the total excitation current  $i'_x$ , as shown in Figure 5.28c. The small-signal voltage on the gate of  $M_2$  is  $-i'_x R_3$  and hence current inject out of  $M_2$ 's drain,  $i_2$ , is  $g_{m2} R_3 i'_x$ . The total current injected into  $R_1$  is  $i'_x + i_2$  which is  $-v_\pi$ , hence  $i'_y = -g_{m1} R_1 (i_1 + i_2) = -g_{m1} R_1 (1 + g_{m2} R_3) i'_x$ , hence

$$T' \equiv -\frac{i'_y}{i'_x} = g_{m1} R_1 \frac{r_{m2}}{r_{m2} + R_3} \quad (5.58)$$

which is again different from the return ratio calculated in (5.54).

Now, if we use (5.34), we have (considering that  $H'_0 = 0$  or equivalently  $D_0 = 1$ ),

$$\begin{aligned} H &= H'_\infty D'_T = H'_\infty \cdot \frac{T'}{1 + T'} \\ &= \frac{g_{m1} g_{m2} R_3 (R_1 + R_2) + g_{m1} R_1}{1 + g_{m1} R_1 (1 + g_{m2} R_3)} \end{aligned}$$

which is exactly the same end result obtained in (5.56).

The last two examples clearly demonstrate that the return ratio is not an invariant of the loop even for a single loop at low frequencies.

#### Example 5.2.6 (Cherry-Hooper Transimpedance Amplifier: Active Feedback)

Consider the Cherry-Hooper transimpedance amplifier shown in Figure 5.29a. Often times the input current source represents the drain current of a transistor and this stage is used as a load at the transistor drain to provide additional peaking to obtain a more broadband response. We will analyze it assuming that it is driven with an ideal current source,  $i_{in}$ .

At low frequencies, we can determine  $H_\infty$  with respect to the transconductance of  $M_1$  by scaling that dependent source up with a factor  $k$  and let  $k \rightarrow \infty$ . In this case, the only way to have a finite output is for the  $v_{gs}$  to approach zero, which in general also implies that the current from that node to ground is also zero. In this case, the input current,  $i_{in}$  flows in its entirety into the  $R_3$  and  $r_{m2}$  in the T-model of the  $M_2$ , as shown in Figure 5.29b. This means that the small-signal voltage at the gate of  $M_2$  is simply  $-i_{in}(R_3 + r_{m2})$ . To have this voltage the output voltage which is related to this voltage via the voltage divider formed between  $R_1$  and  $R_2$  must be  $1 + R_2/R_1$  times the voltage at the gate of  $M_2$ , thus,

$$H_\infty \equiv \frac{v_{out}}{i_{in}}|_{k \rightarrow \infty} = -(R_3 + r_{m2}) \cdot \left(1 + \frac{R_2}{R_1}\right)$$

It is easy to see that  $H_0 = 0$  since  $v_o = 0$  for  $k = 0$ .

The return ratio with respect to the transconductance source of  $M_1$  is easily calculated by using the source  $i_z$ , as shown in Figure 5.29c. Current  $i_x$  flows out of  $R_1$  producing a voltage of  $i_x R_1$  at the gate of  $M_2$ . At low frequencies, this voltage appears directly at the gate of  $M_1$  through the unity gain source-follower formed by  $M_2$ . This voltage will induce an  $i_y = -g_{m1}R_1$ , hence

$$T \equiv -\frac{i_y}{i_x} = g_{m1}R_1$$

At high frequencies we consider only two capacitors, namely,  $C_1$  between the input and ground and  $C_2$  between output and ground. In this case,  $H_\infty$  with respect to the transconductance source of  $M_1$  remains unchanged. This is because  $v_{gs} \rightarrow 0$  would imply no current will flow through  $C_1$ . Also  $C_2$  is in parallel with the output and hence does not change the voltage needed at the output to produce the right voltages and currents elsewhere in the circuit.  $H_0$  is still zero since there is still no path for the signal to reach the output in the absence of the transconductance of  $M_1$ .

On the other hand the return ratio does change at high frequencies. This time  $i_x$  is divided between  $R_1 + R_2$  and  $C_2$ , with a current divider ratio of  $1/[1 + (R_1 + R_2)C_2 s]$ . Also in the presence of  $C_1$  the gain of the source follower formed by  $M_2$  in the feedback path is given by  $1/[1 + (R_3 + r_{m2})C_1 s]$ . Thus, the high frequency return ratio is given as

$$\begin{aligned} T(s) &\equiv -\frac{i_y}{i_x} = g_{m1}R_1 \cdot \frac{1}{[1 + (R_3 + r_{m2})C_1 s][1 + (R_1 + R_2)C_2 s]} \\ &= T(0) \cdot \frac{1}{(1 + \tau_1 s)(1 + \tau_2 s)} \end{aligned} \quad (5.59)$$

where  $T(0) = g_{m1}R_1$ ,  $\tau_1 = (R_3 + r_{m2})C_1$ , and  $\tau_2 = (R_1 + R_2)C_2$ .

$$D_T(s) = \frac{T}{1 + T} = \frac{T(0)}{1 + T(0)} \cdot \frac{1}{1 + \frac{\tau_1 + \tau_2}{1 + T(0)}s + \frac{\tau_1 \tau_2}{1 + T(0)}s^2} \quad (5.60)$$

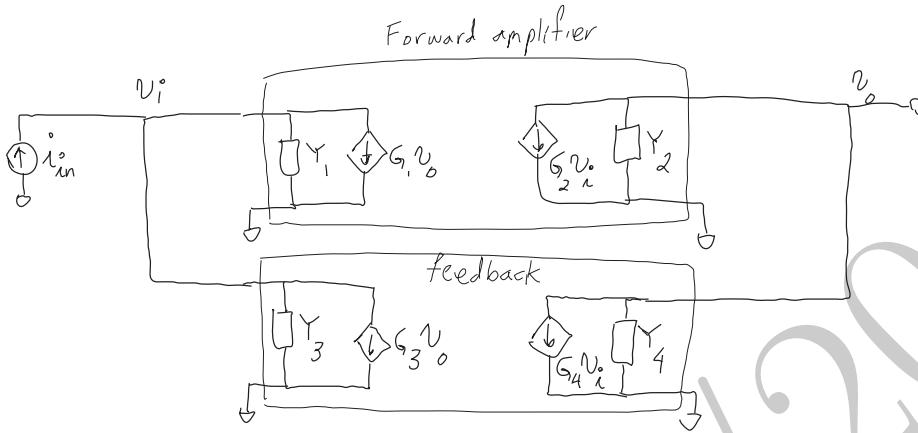


Figure 5.30: Two-port models for the amplifier and the feedback network for bilateral networks.

*It is easy to verify that the poles will become a complex conjugate pair (when  $Q > \frac{1}{2}$  or equivalently  $\zeta < 1$ ) for large values of the return ratio, i.e., for*

$$T(0) > \frac{(\tau_1 - \tau_2)^2}{4\tau_1\tau_2}$$

*This allows us to control the damping ratio of the Cherry-Hooper stage and produce under-damped and peaking responses that are generally obtained with shunt-peaking in much smaller area. The price paid is the more nonlinear response of the stage, particularly single there is a nonlinear stage in the feedback that will get inverted as discussed earlier.*

### 5.2.3 Challenges with Return Ratio

INTERMEDIATE TOPIC

As mentioned earlier, the value of the return ratio depends on the controlled source for which the return ratio is evaluated. An illustrated example is shown in Figure 5.30 consisting of two general two-ports to account for the forward path and the feedback network. The admittances,  $Y_1$  and  $Y_2$  represent the input and output admittances of the forward path (amplifier) and transconductances,  $G_1$  and  $G_2$  capture its reverse and forward transconductances that can be frequency dependent in general. Similarly for the feedback network,  $Y_3$  and  $Y_4$  represent its output and input admittances, respectively, while its forward and reverse transconductances are denoted by  $G_3$  and  $G_4$ . For brevity, we also define  $Y_i = Y_1 + Y_3$ ,  $Y_o = Y_2 + Y_4$ ,  $G_i = G_1 + G_3$ , and  $G_o = G_2 + G_4$ . Note that the desired primary signal flow direction in the feedback network of Figure 5.30 is from right to left.

First, let us evaluate the return ratio with respect to  $G_2$ . To do so, we can place a current source  $i_z$  in parallel with it and calculate the returned current,

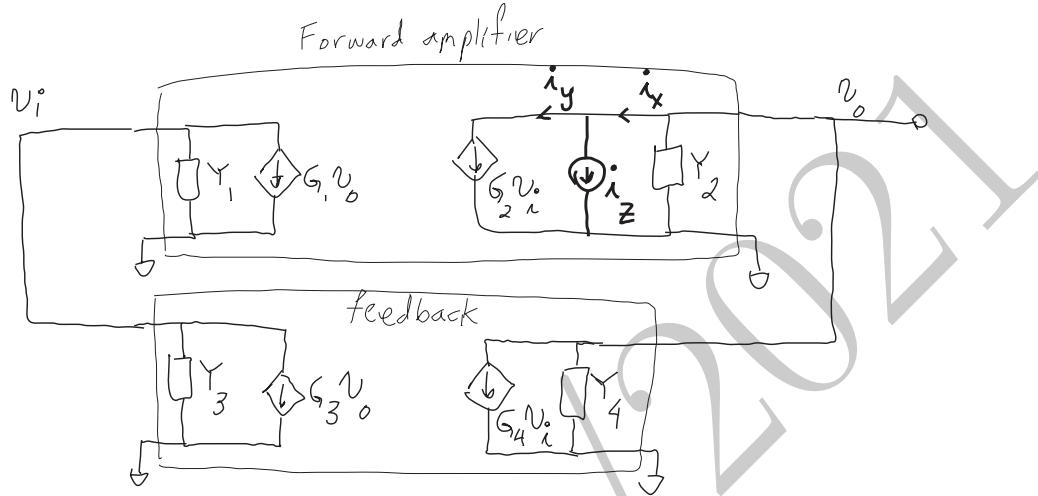


Figure 5.31: The bilateral two-port with test current source to determine the return ratio.

$i_y$ , as a function of the total injected current,  $i_x$ , as shown in Figure 5.31. To calculate the return ratio, we must apply the KCL at both sides of  $i_z$ , as well as on the input side with no input ( $i_{in} = 0$ ),

$$i_x + G_4 v_i + (Y_o) v_o = 0 \quad (5.61a)$$

$$G_2 v_i = i_y \quad (5.61b)$$

$$(G_i) v_o + (Y_i) v_i = 0 \quad (5.61c)$$

which can be solved to find the return ratio of  $G_2$  as<sup>19</sup>,

$$T_{G_2} \equiv -\frac{i_y}{i_x} = -\frac{G_2 G_i}{Y_i Y_o - G_i G_4} = -\frac{G_2 G_3 + G_1 G_2}{Y_i Y_o - G_1 G_4 - G_3 G_4} \quad (5.62)$$

Next we determine the return ratio associated with the dependent source  $G_3$  which is also applied in the forward direction of the loop (clockwise in this case). In a similar fashion, we can write calculate the return ratio of  $G_3$  to be

$$T_{G_3} = -\frac{G_3 G_o}{Y_i Y_o - G_1 G_o} = -\frac{G_2 G_3 + G_3 G_4}{Y_i Y_o - G_1 G_4 - G_1 G_2} \quad (5.63)$$

which is clearly different from  $T_{G_2}$ , the return ratio of  $G_2$  calculated in (5.62). We see that these quantities are the same only when both forward and reverse paths are unilateral, i.e.,  $G_1 = G_4 = 0$ , in which case they both reduce to

$$T_{uni} = -\frac{G_2 G_3}{Y_i Y_o} \quad (5.64)$$

<sup>19</sup>This result could be obtain in a more straightforward manner from the general definition of the return ration given by (5.66) later in subsection 5.2.4

This shows that the return ratio is not an invariant of the loops, even for a single loop at low frequencies, unless both forward and reverse parts are unilateral. We will introduce the loop gain later, which does not suffer from this limitation.

#### 5.2.4 General Definition of Return Ratio

In the section 5.2, we defined the returned ratio for a dependent source,  $g$ , (similar to that of Figure 5.22) as the ratio of the signal returned by the controlled source, namely  $u_y$ , to the independent source,  $u_x$ , that has replaced it in the circuit. This is a useful definition if we try to calculate the return ratio of a given circuit.

There is another way to define the return ratio for a given controlled source in an arbitrary circuit<sup>20</sup> in terms of the determinant of the circuit defined in (2.19) in subsection 2.2.1 of Chapter 2.

The *return difference* of source  $g$  is defined as one plus the return ratio  $(1 + T_g)$  and is given by the ratio of the determinant of the  $Y$ -matrix of the circuit under normal circumstances,  $\Delta$ , divided by the determinant of the circuit when the reference dependent source disappears ( $g = 0$ ), shown as,  $\Delta_g$ , i.e.,

$$1 + T_g \equiv \frac{\Delta}{\Delta_g} \quad (5.65)$$

or equivalently

$$T_g \equiv \frac{\Delta}{\Delta_g} - 1 \quad (5.66)$$

For example, it can be easily verified that (5.66) produces the same result as (5.62) for the return ratio of  $G_2$  in the general bilateral two-port model of Figure 5.30. The above definition of return ratio can be arrived at from the basic definition based on a controlled source given in subsection 5.2.1. To see how this is done for a circuit in general, see [?, page 46].

### 5.3 Effect of Feedback on Impedance Levels

Online YouTube lecture:

[Port Impedance under Feedback, Blackman Formula](#)

#### 5.3.1 Blackman's Formula

As we saw in chapter 4, the impedance seen looking into a port of a circuit is yet another transfer function where the input variable  $u_i$  is  $i_t$  and the output variable  $u_o$  is the induced voltage at the *same* port, namely  $v_t$ . The impedance is thus defined as  $Z \equiv v_t/i_t$ . We can use the formulation of (5.29) to express the impedance in terms of some of the easily measurable parameters of the circuit. ▼ Derivation ▼

<sup>20</sup>This definition is due to Bode and can be found on p. 49 of [Bode's book].

First we determine  $Z_0$  which is the impedance seen at the port of interest for  $k = 0$  and can be thought of as *direct transmission* similar to  $H_0$ . From (5.31) it is obvious that

$$Z_0 \equiv \frac{v_t}{i_t}|_{k=0} = A \quad (5.67)$$

Now we determine the return ratio similar to the standard  $T$  when the input current is zero ( $i_t = 0$ ). However, in this case,  $i_t = 0$  corresponds to the open circuited input, hence, we refer to it as the *open circuit* return ratio,  $T_{open}$ . It can be easily calculated from (5.29) by setting  $u_i$  (i.e.,  $i_t$ ) to zero. In this case, (5.29a) reduces to  $u_o = Bu_x$  and (5.29b) becomes  $u_a = Du_x$ , which combined with the amplification of the dependent source ( $u_y = gu_a$  under nominal gain condition, i.e.,  $k = 1$ ), simply result in

$$T_{open} \equiv -\frac{u_y}{u_x}|_{i_t=0} = -gD \quad (5.68)$$

Last, we determine the loop gain when the output variable  $v_t$  is zero, which is equivalent to the port being short circuited (hence  $v_t = 0$ ) under nominal gain condition ( $k = 1$ ). The return ratio with the port shorted,  $T_{short}$ , is determined by setting  $u_o$  to zero in (5.29) and eliminating  $u_i$  between the two resulting equations that leads to

$$T_{short} \equiv -\frac{u_y}{u_x}|_{v_t=0} = -\frac{gu_a}{u_x} = g\left(\frac{BC}{A} - D\right) \quad (5.69)$$

We have already determined the general transfer function in (5.31) which is stated in terms of the  $A$ ,  $B$ ,  $C$ , and  $D$  parameters of (5.29). We can reorder the result of (5.29) so that we can express it in terms of the parameters  $Z_0$ ,  $T_{open}$ , and  $T_{short}$ , specified in (5.67), (5.68), and (5.69), respectively. Under nominal gain condition ( $k = 1$ ) it reduces to:

$$Z \equiv \frac{v_t}{i_t} = A + \frac{gBC}{1-gD} = A \cdot \frac{1+g\left(\frac{BC}{A}-D\right)}{1-gD} \quad (5.70)$$

### ▼ Result ▼

This directly leads to Blackman's impedance formula which expresses the impedance seen looking into any desired port of the circuit as

$$Z = Z_0 \cdot \frac{1 + T_{short}}{1 + T_{open}}$$

(5.71)

where  $Z_0$  is the impedance seen looking into the same port with the reference dependent source going to zero ( $k = 0$ ), and  $T_{open}$  and  $T_{short}$  are the return ratios calculated with respect to the *same* dependent source ( $ku_a$ ), when the port of interest is open- and short-circuited, respectively. We will refer to  $T_{open}$  and  $T_{short}$  as *open-port* and *short-port* return ratios, respectively. The Blackman's formula is particularly useful when the reference source is in the feedback path and disables the feedback loop, making the calculation of  $Z_0$  easy.

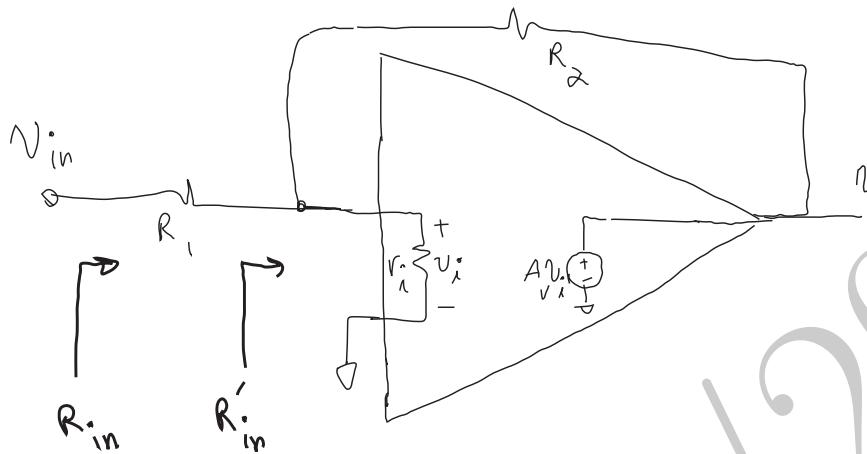


Figure 5.32: An op-amp in inverting configuration with finite input and non-zero output resistance.

### Determination of the feedback type

In fact Blackman's formula can be used to determine the "type" of the feedback configuration. In general, we refer to a configuration as *shunt* (at the input and/or the output) when  $T_{open} > T_{short}$  and as *series* when  $T_{open} < T_{short}$ . This is perhaps the most straight forward way to determine the configuration of a feedback network.

Also it is noteworthy that the feedback configuration is indeed a function of the input and output terminals. If the input is applied to a different terminal of exactly the same circuit or the output is taken from a different terminal of the same circuit, the configuration will be different. In fact, sometime in the same circuit there are two output nodes, the output impedance of which goes in opposite direction with application of feedback (i.e., one increases while the other decreases). In this case one of the output terminals is in series configuration while the other one is in shunt arrangement. An instance of this is discussed later in Example 5.3.3. It is feedback stage of Example 5.1.4, shown in Figure 5.6, where the regular output taken from the drain of  $M_2$  is in series configuration and thus increases, while the output taken from the source of  $M_2$  is in a shunt arrangement and hence is reduced by the feedback.

**Example 5.3.1 (Op-Amp with finite input resistance)** *In this example, we will revisit the inverting op-amp configuration of example 5.1.9 without the capacitor in the feedback path. This time we assume a finite voltage gain,  $A_v$ , for the op-amp as well as a finite input impedance,  $r_i$ , as depicted in Figure 5.32. First, let us determine the input resistance. Scaling the dependent voltage source gain to  $kA_v$ , we can calculate the zero-value resistance,*

$$R_{in,0} \equiv R_{in}|_{k=0} = R_1 + r_i \parallel R_2 \quad (5.72)$$

which is obtained noting that for  $k = 0$  the dependent voltage source is simply a short-circuit. Now we can determine the nominal-gain ( $k=1$ ) dependent voltage source return ratios for the cases when the input is short and open circuited. For short circuited input we have,

$$T_{short} = A_v \cdot \frac{r_i \parallel R_1}{R_2 + r_i \parallel R_1} \quad (5.73)$$

and for an open circuited input,  $r_i \parallel R_1$  becomes  $r_i$ , i.e.,

$$T_{open} = A_v \cdot \frac{r_i}{r_i + R_2} \quad (5.74)$$

Therefore the input impedance is given by Blackman's formula, (5.71), as:

$$\begin{aligned} R_{in} &= R_{in,0} \cdot \frac{1 + T_{short}}{1 + T_{open}} = (R_1 + r_i \parallel R_2) \cdot \frac{1 + A_v \frac{r_i \parallel R_1}{r_i \parallel R_1 + R_2}}{1 + A_v \frac{r_i}{r_i + R_2}} \\ &= R_1 + \frac{r_i R_2}{R_2 + (1 + A_v)r_i} \end{aligned} \quad (5.75)$$

It is noteworthy that for large op-amp gain, the second term in the input impedance diminishes to zero and hence the input impedance approaches  $R_1$ . This could be concluded from the asymptotical equality principle, since that implies that voltage of the inverting op-amp input is asymptotically equal to its other terminal which is at ground, hence the impedance seen on the input node of the op-amp will be asymptotically zero for large  $A_v$ .

We can see that in this case, both  $T_{short}$  and  $T_{open}$  are non-zero. Although, the Blackman equation always provide us with the correct result, we can simplify our calculation by "preprocessing" the circuit a little bit. Looking back at Figure 5.32, we notice that  $R_{in}$  consists of  $R_1$  is in series the impedance seen between the input node and ground that we denote as  $R'_{in}$ . It is easier to apply (5.71) to  $R'_{in}$ , since shorting that node to ground will render the return ratio zero, i.e.,  $T'_{short} = 0$ . The zero-value resistance is given by

$$R'_{in,0} \equiv R'_{in}|_{A_v=0} = r_i \parallel R_2 \quad (5.76)$$

The open circuit return ratio is simply

$$T'_{open} = A_v \cdot \frac{r_i}{r_i + R_2} \quad (5.77)$$

thus,

$$R'_{in} = R'_{in,0} \cdot \frac{1}{1 + T'_{open}} = \frac{r_i \parallel R_2}{1 + A_v \frac{r_i}{r_i + R_2}} = \frac{r_i R_2}{R_2 + (1 + A_v)r_i} \quad (5.78)$$

which demonstrates the asymptotic short circuit to ground more directly as  $A_v$  become very large. Now it is obvious that

$$R_{in} = R_1 + R'_{in} \quad (5.79)$$

**Example 5.3.2 (MOS Shunt-Shunt Amplifier:Input and Output Resistance)**

In Examples 5.1.5 and 5.2.2, we studied the shunt-shunt stage, shown in Figure 5.7. Now we can apply the Blackman's formula in (5.71) to determine its input and output resistances. Let us start with the input impedance.

To be able to reuse the return ratio obtained in Example 5.2.2, we should determine the  $R_0$  (or  $Z_0$  in (5.71) generally speaking) for with respect to the  $T$ -model dependent current source of  $M_2$  when  $k = 0$ . In that case, assuming that  $r_{o1} \gg r_{m2}$ , the input  $R_0$  is simply the resistance seen looking into the source of  $M_2$ . Ignoring body-effect it is

$$R_{in,0} \equiv R_{in}|_{k=0} = r_{m2} \quad (5.80)$$

The open-port return ratio,  $T_{open}$ , with respect to the same dependent source is the same as the return ratio calculated in (5.47) in Example 5.2.2, since nulling the input current source of Figure 5.7 corresponds to opening of the input.

The return ratio for the shorted port,  $T_{short}$  with respect to the same source can be easily seen to be zero, since a short circuit at the input will sink all of the small-signal drain current of  $M_1$  and hence the returned current at the drain of  $M_2$  will be zero.

Now combining these results we obtain the input resistance from (5.71) as

$$R_{in} = r_{m2} \cdot \frac{1}{1 + T} \quad (5.81)$$

where  $T$  is given by (5.47).

The output resistance can be obtained in a similar fashion using (5.71). First to determine  $Z_0$ , we set  $k$  to zero and determine the impedance seen looking into the output terminal, which is the parallel combination of the load resistor,  $R_3$  and the feedback network,  $R_1 + R_2$ ,

$$R_{out,0} \equiv R_{out}|_{k=0} = R_3 \parallel (R_1 + R_2) \quad (5.82)$$

Again, the return ratio with reference to  $k$  with the output port opened,  $T_{open}$ , is the same as the one calculated in (5.47). Also in this case the port-short return ratio,  $T_{short} = 0$  because a short to ac ground at the output forces the gate voltage of  $M_1$  to be zero. Thus,

$$R_{out} = [R_3 \parallel (R_1 + R_2)] \cdot \frac{1}{1 + T} \quad (5.83)$$

with  $T$  given by (5.47).

Note that this shunt-shunt feedback stage reduces both the input and the output impedances.

It is easy to see that in this configuration at both the input and the output ports,  $T_{open} > T_{short}$ , showing that we have a shunt-shunt configuration.

**Example 5.3.3 (Output Resistance of MOS Shunt-Series Feedback Stage)**

Let us revisit the MOS shunt-series feedback stage of Example 5.1.4, shown in

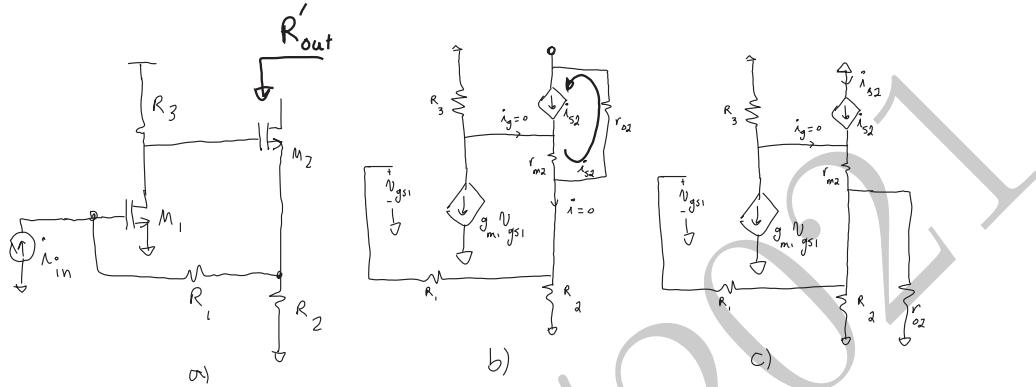


Figure 5.33: a) the intrinsic output resistance of the shunt-series MOS stage of Figure 5.6, b) calculation of  $T_{open}$  where the drain current circulates inside the transistor (through  $r_{o2}$ ), c) calculation of  $T_{short}$ .

*Figure 5.6a.* We are interested in the intrinsic output resistance,  $R'_{out}$ , which in parallel with  $R_4$ , determines the total output resistance, as shown in Figure 5.33a.

However, this time if we ignore the output resistance of  $M_2$  (i.e.,  $r_{o2} \rightarrow \infty$ ), this resistance would simply be infinity. Thus, we must take  $r_{o2}$  into account<sup>21</sup>. To apply Blackman's formula we first determine the zero gain ( $k = 0$ ) output resistance,  $R'_{out,0}$  with respect to the  $\pi$ -model dependent current source of  $M_1$ . When  $k = 0$  is zero,  $M_2$  reduces to a common-source stage with  $R_2$  as the source resistance, hence its impedance is simply given by (3.50) (ignoring body-effect) from Chapter 3, i.e.,

$$R'_{out,0} \equiv R'_{out}|_{k=0} = r_{o2}(1 + g_{m2}R_2) \quad (5.84)$$

This time,  $T_{open}$ , the open-port return ratio with respect to the same dependent source is zero. This can be seen by looking at the small-signal equivalent model of the output shown in Figure 5.33b. If the output is opened (assuming that  $R_4$  is already out of the picture), the entire current of the drain current source will flow through  $r_{o2}$ . However, this current is equal to the source current,  $i_{s2}$ , therefore, the current injected into  $R_2$  by the transistor is zero, thus no signal is returned, therefore,  $T_{open} = 0$ .

To determine  $T_{short}$ , we need to short circuit the output. This will simply reduce the  $M_2$  to a source follower stage. In this case, all of the drain current of  $M_2$  will flow through the short circuit, and no current flows into  $r_{o2}$ , as shown in Figure 5.33c. In this case,  $r_{o2}$  is in parallel with  $R_2$  and  $T_{short}$  with respect

<sup>21</sup>The output resistance of  $M_1$ , i.e.,  $r_{o1}$ , has already been taken into account effectively, since it can be simply absorbed into  $R_3$  which is in parallel with it.

to the controlled source  $g_{m1}$  in a similar fashion as in Example 5.2.1 to be

$$T_{short} = g_{m1}R_3 \cdot \frac{R_2 \parallel r_{o2}}{r_{m2} + R_2 \parallel r_{o2}} \quad (5.85)$$

which for  $R_2 \ll r_{o2}$  reduces to

$$T_{short} \approx g_{m1}R_3 \cdot \frac{R_2}{r_{m2} + R_2} \quad (5.86)$$

which is the same as the return ratio calculated in (5.41) of Example 5.2.1, where we ignore  $r_{o2}$  to begin with.

which allows us to calculate the intrinsic output resistance using (5.71) to be

$$R'_{out} = R'_{out,0}(1 + T_{short}) \approx r_{o2}[1 + g_{m2}R_2(1 + g_{m1}R_3)] \quad (5.87)$$

Now let us verify this in a numerical example. Assume a transconductance of  $g_m = 20mS$  for both  $M_1$  and  $M_2$  and an  $r_{o2}$  of  $1k\Omega$  for  $M_2$ . Let the feedback resistors be  $R_1 = 900\Omega$  and  $R_2 = 100\Omega$ , while the drain resistance of  $M_1$  is  $R_3 = 5k\Omega$ . Assuming a large  $R_4$ , (5.87) predicts  $R'_{out} = 203k\Omega$ . An AC analysis in SPICE predicts a low frequency output impedance of  $R'_{out} = 203k\Omega$  in agreement with the analytical result.

Figure 5.34 shows the simulated output impedance as a function of frequency assuming  $C_{gs} = 50fF$  and  $C_{gd} = 20fF$  for both  $M_1$  and  $M_2$ .

As we can see in the case of shunt-series feedback the output impedance increases of the output node taken from the drain of  $M_2$ . In fact, since  $T_{short} > T_{open}$  for the output, we can easily conclude that this output network forms a series combination. However, as discussed earlier, the feedback “type” is a function of both the circuit and the input and output terminals. In the same circuit, if the output were to be taken from the source of  $M_2$  instead, it is easy to see that  $T_{short} = 0$  since shorting that node results in no returned signal and hence,  $T_{short} < T_{open}$  which means that for the output taken from the source of  $M_2$ , the output network is in a series configuration. This is an example of our statement that the same circuit can be in different feedback configurations depending on where the output is taken or the input is applied.

**Example 5.3.4 (Input Impedance of the BJT shunt-shunt stage)** Let us consider the effect of the base-collector capacitance,  $C_\mu$ , in the BJT shunt-shunt amplifier of Example 5.1.3. The  $\pi$ -model dependent current source of the transistor is the natural reference for calculation of the return ratio and zero-value impedance. For  $k = 0$ , we have,

$$Z_{in,0}(s) \equiv Z_{in}(s)|_{k=0} = r_\pi \parallel [Z_\mu(s) + R_3] \quad (5.88)$$

where  $Z_\mu(s) = R_2/(1 + R_2C_\mu s)$  is the impedance of the parallel combination of the  $R_2$  and  $C_\mu$ . It is obvious that the short-port return ratio,  $T_{short}$  is zero, since  $v_\pi$  will be zero if the input is short circuited. The open-port return ratio,

#### ♦ Numerical Example ♦

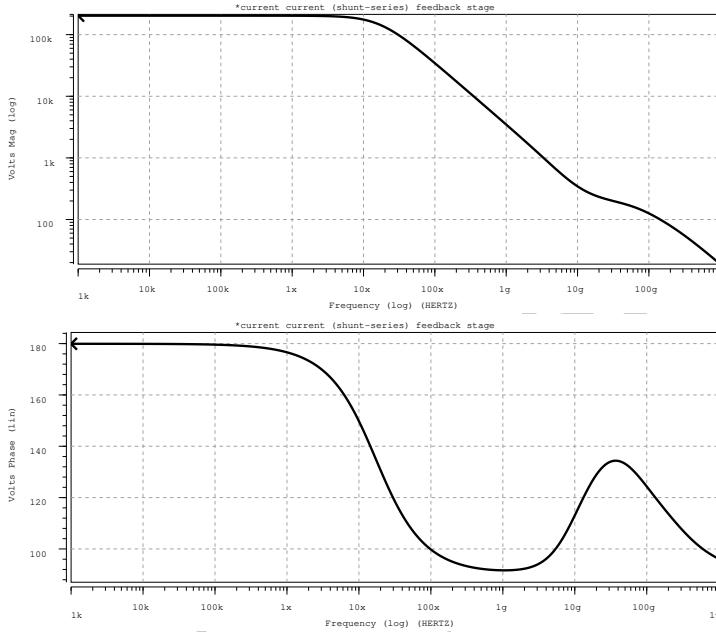


Figure 5.34: The output impedance of the shunt-series feedback MOS amplifier.

is determined by the current divider ratio between  $R_3$  and  $Z_\mu(s) + r_\pi$ , times transistor current gain,  $\beta$ , i.e.,

$$T_{open}(s) = \frac{\beta R_3}{r_\pi + R_3 + Z_\mu(s)} \quad (5.89)$$

which predicts an input impedance of

$$\begin{aligned} Z_{in}(s) &= \frac{r_\pi [Z_\mu(s) + R_3]}{r_\pi + (1 + \beta)R_3 + Z_\mu(s)} \\ &= \frac{r_\pi (R_2 + R_3)}{r_\pi + R_2 + (1 + \beta)R_3} \cdot \frac{1 + (R_2 \parallel R_3)C_\mu s}{1 + R_2 \parallel [r_\pi + (1 + \beta)R_3]C_\mu s} \\ &= R_{in} \cdot \frac{1 + \tau_z s}{1 + \tau_p s} \end{aligned} \quad (5.90)$$

where  $R_{in}$  is the low frequency input resistance which is smaller than the open-loop input resistance,  $r_\pi$ .

♦ Numerical Example ♦

Let us look at a numerical example with transconductance of  $g_m = 40mS$ , a transistor  $\beta$  of 100, and  $C_\mu = 100fF$ . The resistors are  $R_2 = 10k\Omega$  and  $R_3 = 2k\Omega$ . In this case, we have  $r_\pi = 2.5k\Omega$ , for which we can calculate the low frequency input resistance of  $R_{in} = 140\Omega$  much smaller than  $r_\pi$ . From (5.90),

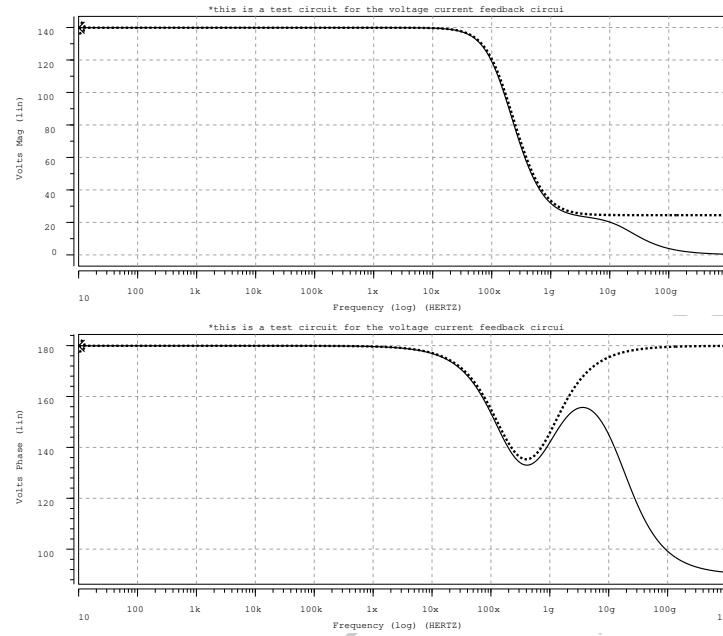


Figure 5.35: The input impedance of the shunt-shunt feedback BJT amplifier with (dotted) and without (solid)  $C_\pi$ .

the pole and zero time constants can be easily calculated to be  $\tau_z = 167\text{ps}$  and  $\tau_p = 953\text{ps}$ , for which we have,  $z = -2\pi \cdot 953\text{MHz}$  and  $p = -2\pi \cdot 167\text{MHz}$ .

A SPICE simulation results for the input impedance is shown in Figure 5.35, with  $C_\pi = 0$  (which is what we calculated) and with  $C_\pi = 400\text{fF}$  (more realistic). It predicts a low frequency input impedance of  $R_{in} = 140\Omega$  in close agreement with the analytical result. Also you can see that we have a pole around 170MHz and a LHP zero around 950MHz again in agreement with the predictions. The introduction of  $C_\pi$  introduces a high frequency pole due to the lowered input impedance of the stage, reducing the time constant.

## 5.4 Loop Gain

Online YouTube lecture:

[Unilateral Loop gain, voltage and current loop gains](#)

Although the terms return ratio and loop gain are sometimes used synonymously, they are not generally the same. We saw in Section 5.2 that return ratio is defined with reference to a controlled source. However, as we saw earlier in examples 5.2.4 and 5.2.5 as well as in subsection 5.2.3, the value of the return

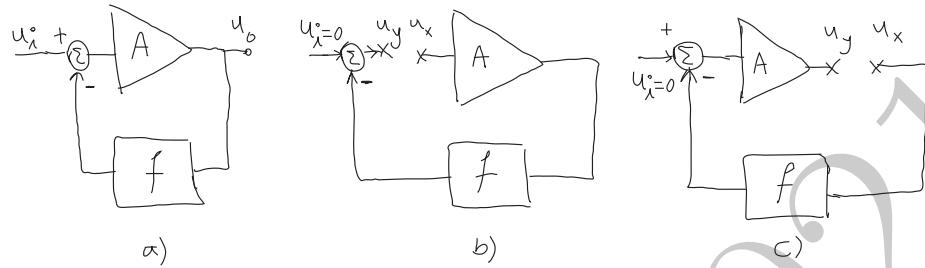


Figure 5.36: a) An idealized feedback loop. Opening of an ideal loop on b) the input side, and c) on the output side.

ratio can vary with the choice of the reference source and in general, can be different if measured with reference to a different controlled sources.

Another problem with calculating the return ratio is that it can only be determined by applying a test source directly in parallel with a controlled current source or in series with a dependent voltage source before any other element. While in some examples it may be easy to find and access such a source, in general, it can be quite challenging to access the independent source before any internal parasitic elements of the transistor, such as the parasitic capacitors or resistors. This can be particularly inconvenient when using compact transistor models in simulators that generate and maintain these parasitic elements internally and automatically.

The lack of invariance of the return ratio of the measurement reference and the associated practical challenge in determining it brings us to another quantity which does not suffer from these shortcomings called the *loop gain*. Loop gain can be determined irrespective of a reference dependent source and is the more appropriate parameter for determination of the stability of the system. In this section we first look at loop gain in a unilateral loop and then generalize it to a bilateral one.

### Ideal Loop Gain

Let us consider the most idealized feedback amplifier model of Figure 5.36a. First let us assume that both the amplifier and the feedback network are unilateral, in other words there is no gain from their outputs back to their inputs<sup>22</sup>. We will deal with the bilateral case in subsection 5.4.2.

If we view the input and output variables as abstract signals (not voltages and currents for which impedance levels do not matter), we can break the loop on the input or the output side, as shown in Figures 5.36b and 5.36c. If we apply a test signal,  $u_x$ , the signal on the other side of the broken circuit,  $u_y$ , is  $-Af u_x$  no matter where we break the loop. So the signal experiences a gain of

<sup>22</sup>Loosely speaking we could say that in this case the signal only flows in the clockwise direction.

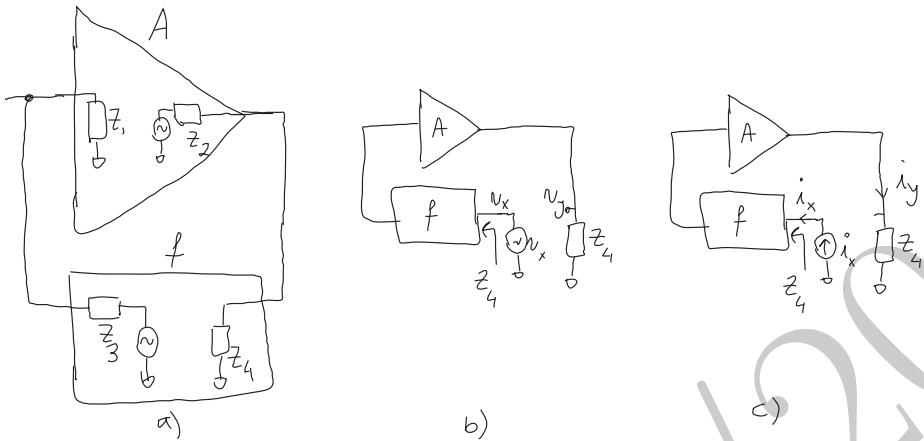


Figure 5.37: Opening of a the feedback loop and replacing the equivalent impedance on the output side. The same loop gain can be measured using either a) test voltage, or b) a test current.

$-Af$  around the loop irrespective of the breaking point. The minus sign appears due to the negative feedback and the quantity  $Af$  is commonly referred to as the *loop gain* although it really is *negative* loop gain. We maintain the same naming convention to stay consistent with this almost universal, yet potentially misleading, naming convention by ignoring the minus sign. The invariance of the loop gain with the measurement point for this idealized unilateral loop is a useful property that would be great to retain for a generally bilateral circuit with finite impedance levels. We will first deal with the finite impedance levels in a unilateral loop and then generalize it to a bilateral one in subsection 5.4.2.

### 5.4.1 Loop Gain of a Unilateral Loop

Although the definition of the loop gain seems straightforward at the block diagram level, there are important nuances when we try to calculate it at the circuit level. This problem arises from the fact that there is a finite impedance on almost every point of a practical circuit. The question is what kind of a source (e.g., voltage or current) and what load impedance on the other side of the broken loop should be used to obtain a loop gain *invariant* of the measurement point.

Now let us assume that both the amplifier and the feedback networks have finite input and output resistances, as depicted in Figure 5.37a. If we break the loop at the output for instance, we want to make sure that the output of the amplifier,  $A$ , still “sees” the same impedance as it did when it was connected to the input of the feedback network. This can be done by placing a replica of the input impedance of the feedback network on the output of the amplifier.

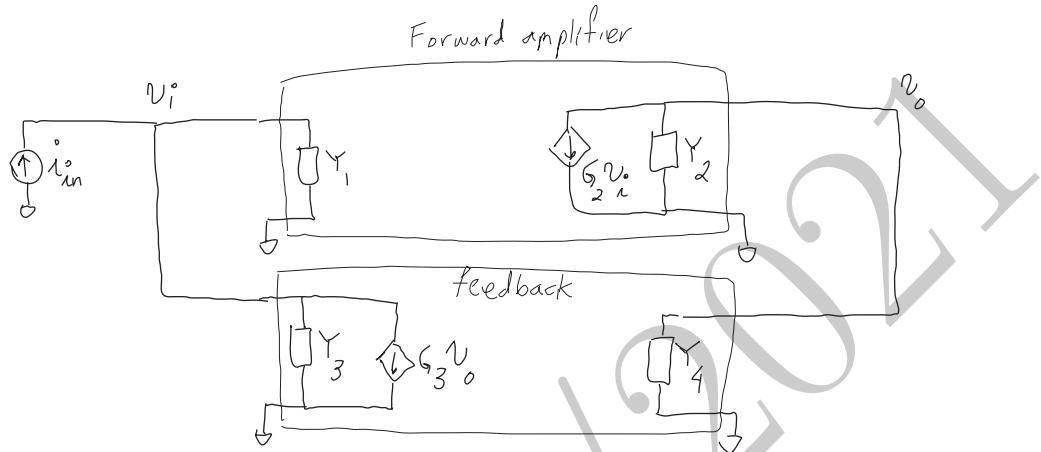


Figure 5.38: Two-port models for the amplifier and the feedback network for a unilateral loop.

Then we can measure the loop gain by applying an ideal<sup>23</sup> test voltage  $v_x$  and measuring the induced voltage  $v_y$  as in Figure 5.37b or a test current  $i_x$  and measure the current  $i_y$  similar to 5.37c, which will produce the same result for the loop gain,

$$T \equiv -\frac{v_y}{v_x} = -\frac{i_y}{i_x} \quad (5.91)$$

To see how this works in a circuit, consider the unilateral feedback system of Figure 5.38 where the amplifier and the feedback network are both modeled as unilateral two-ports<sup>24</sup>. Breaking the loop on the output side, we can drive the input of the feedback network with a voltage source,  $v_x$ . This results in a voltage,

$$v_i = -v_x \frac{G_3}{Y_i}$$

on the input side ( $Y_i = Y_1 + Y_2$ ), which in turn results in an output voltage,

$$v_y = v_x \frac{G_2 G_3}{Y_i Y_o}$$

where  $Y_o = Y_2 + Y_4$ . Hence according to (5.91), the loop gain of this unilateral

<sup>23</sup>One might wonder why we do not use a voltage source with a source resistance  $Z_2$  to drive the input of the feedback network. The answer can be found in the substitution theorem that states we can replace a branch with the voltage it experiences without changing the voltages and currents of the other branches and nodes. In this case, in effect we are replacing the voltage seen between  $Z_2$  and  $Z_4$  with the voltage source. This voltage is determined by the voltage divider formed by  $Z_2$  and  $Z_4$  the same as the closed loop case.

<sup>24</sup>This model could be obtained from the complete two-port model of Figure 5.30 by setting  $G_1$  and  $G_4$  to zero.

loop can be determined as negative of the ratio of  $v_y$  to  $v_x$ ,

$$\mathbb{T}_{uni} = -\frac{G_2 G_3}{Y_i Y_o} \quad (5.92)$$

Similarly, we could inject a current source  $i_x$  and measure the current through the admittance,  $Y_4$ , namely,  $i_y$ , to determine the loop gain based on (5.91). It is easy to verify that calculating the loop gain as the negative of the ratio of  $i_y$  to  $i_x$  will result in exactly the same result as (5.92).

It is possible to see that breaking this unilateral loop at any other point (either on the input or the output, before or after the source admittance) and using a similar procedure would result in the same result as (5.92). Producing an invariant loop gain at least for a unilateral loop.

An important observation can be made here. Comparing (5.64) and (5.92), we notice that in the case of a unilateral loop the return ratio calculated for any dependent source in the forward direction of the loop (clockwise in our example) and the invariant loop gain are indeed equal. This is a useful result if we already know we are dealing with a unilateral loop which allows us to substitute loop gain and return ratio for each other. Depending on the case, one may be easier to calculate (or measure) and hence simplify our analysis. However, one should be *very careful* not to over-generalize this special result that only applies to a unilateral loop. This has been a source of prolonged confusion in the literature and should be avoided. Also keep in mind that almost all practical circuits are bilateral at higher frequencies due to the parasitic elements, such as capacitors, that create bilateral signal paths in the circuit.

### Determination of the Forward Loop Gain with Two Measurements

Although the above procedure can always be applied to determine the loop gain by placing the appropriate *loading* impedance on the second side of the opened loop, it required us to determine what this impedance actually is. This can be somewhat challenging particularly at higher frequencies due to the frequency dependence of the impedances. We can instead obtain the same result by performing two separate measurements of the voltage and current gains, as described next.

The loop can be broken anywhere. Figure 5.39 shows the case, where it is broken on the output side. The forward *voltage* loop gain is measured with an ideal voltage source,  $v_x$ , applied in the forward loop direction (clockwise in Figure 5.39) and the returned voltage measured with the other broken end *open-circuited* ( $i_y = 0$ ), as shown in Figure 5.39a. The forward voltage loop

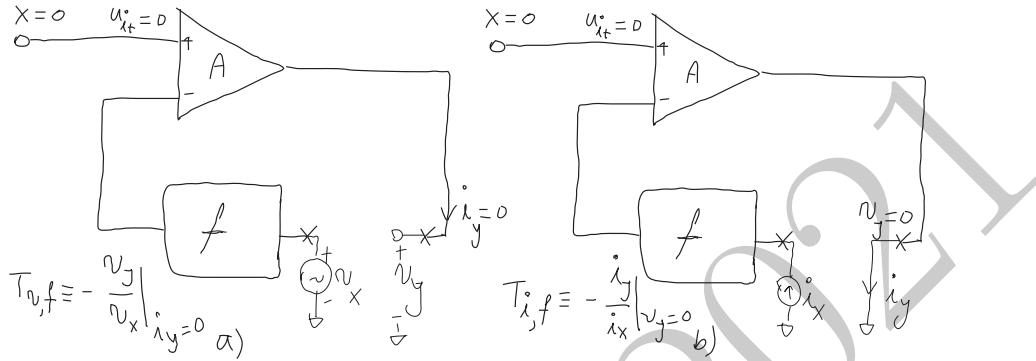


Figure 5.39: Breaking the loop for calculation of the forward a)voltage loop gain, and b)current loop gain.

gain<sup>25</sup>,  $\mathbb{T}_v$ , is hence defined as<sup>26</sup>

$$\mathbb{T}_v \equiv -\frac{v_y}{v_x}|_{i_y=0} \quad (5.93)$$

Note that the minus sign is to account for the fact that for negative feedback,  $\mathbb{T}_v$  has the opposite sign as the gain around the loop.

Next a test *current* source,  $i_x$ , is applied to the input of the feedback network while the other side of the broken loop is *short-circuited* ( $v_y = 0$ ) with the returned current being called,  $i_y$ , as shown in Figure 5.39b. The forward current loop gain,  $\mathbb{T}_i$ , which always measured with the other end *shorted*, is defined as,

$$\mathbb{T}_i \equiv -\frac{i_y}{i_x}|_{v_y=0} \quad (5.94)$$

### ▼ Derivation ▼

At the circuit level, the case of a unilateral loop with an arbitrary main amplification path can be visualized using two-port networks shown in Figure 5.38.

We can open the loop on the input or the output side. We have shown the loop being opened on the output side, in Figure 5.40. To measure the forward voltage loop gain,  $\mathbb{T}_v$ , we apply a test voltage  $v_x$  to the input of the feedback network and measured the voltage,  $v_y$ , on the other broken end which is *open-circuited*. Applying the KCL at the input and output nodes, we obtain:

$$G_2 v_i + Y_2 v_y = 0 \quad (5.95a)$$

$$Y_i v_i + G_3 v_x = 0 \quad (5.95b)$$

<sup>25</sup>We will not use an index *f* to distinguish the forward loop gain unless the *reverse* loop gain is also important. So any loop gain with no reference to forward or reverse is a forward loop gain.

<sup>26</sup>By  $\mathbb{T}_v$  we really mean,  $\mathbb{T}_{v,open}$  which emphasizes that this measurement is performed with the other side open. However, for notational brevity we drift the ‘open’ (and ‘short’ for the current) subscripts. So it should be kept in mind that whenever we refer to voltage and current loop gains, the assumption is the other end is open- and short-circuited, respectively.

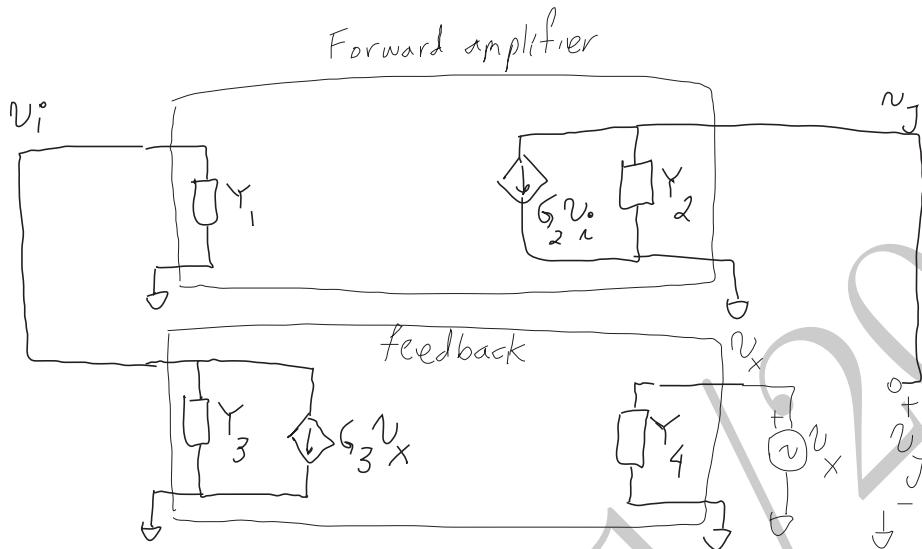


Figure 5.40: Measurement of the forward voltage loop gain,  $\mathbb{T}_v$  in the unilateral feedback loop of Figure 5.38.

The above equations can be easily solved for  $\mathbb{T}_v$  by eliminating  $v_1$  to produce:

$$\mathbb{T}_v \equiv -\frac{v_y}{v_x} \Big|_{i_y=0} = -\frac{G_2 G_3}{Y_i Y_2} \quad (5.96)$$

Now let us calculate the forward current loop gain,  $\mathbb{T}_i$ . This time we apply an ideal current source,  $i_x$ , at the *same point* as the voltage source, i.e., the input of the feedback network and determine the current through the *short circuited* other end of the broken loop, as in Figure 5.41. Again, by applying the KCL to the three independent nodes of the circuit we obtain:

$$Y_i v_1 = i_x \quad (5.97a)$$

$$G_2 v_i = -i_y \quad (5.97b)$$

$$Y_i v_i + G_3 v_1 = 0 \quad (5.97c)$$

which again can be solved for  $\mathbb{T}_i$  to provide,

$$\mathbb{T}_i \equiv -\frac{i_y}{i_x} \Big|_{v_y=0} = -\frac{G_2 G_3}{Y_i Y_4} \quad (5.98)$$

Now we notice that the sum of the reciprocals of the forward current and voltage loop gains,

$$\frac{1}{\mathbb{T}_v} + \frac{1}{\mathbb{T}_i} = -\frac{Y_i Y_o}{G_2 G_3} = \frac{1}{\mathbb{T}} \quad (5.99)$$

is exactly equal to reciprocal of the forward loop gain calculated in (5.92). ▼ Result ▼

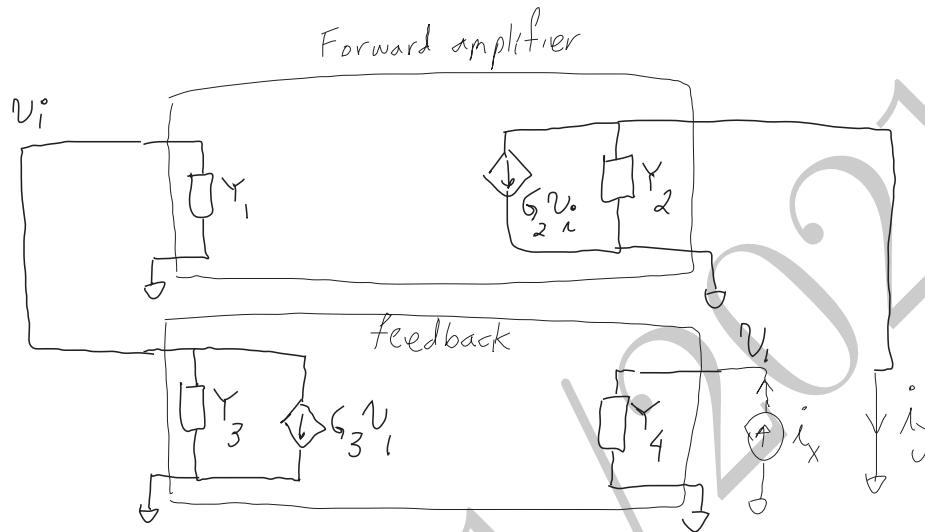


Figure 5.41: Forward current loop gain calculation for the unilateral two-port model of Figure 5.38.

Therefore, we conclude that for the *unilateral* feedback loop, the *forward* loop gain,  $\mathbb{T}$  is related to the forward current and voltage loop gains by<sup>27</sup>:

$$\frac{1}{\mathbb{T}} = \frac{1}{\mathbb{T}_v} + \frac{1}{\mathbb{T}_i} \quad (5.100)$$

In a unilateral loop (where the signal can circulate only in one direction), the loop gain of (5.100) is invariant of where it is measured. This is a useful property for the loop gain as unlike the return ratio its measurement is not controlled by where and how it is done which allows us to pick the most convenient place to break the loop and measure it. We will see in Section 5.4.2 how it can be generalized to the case of a bilateral loop. We will also see there that although (5.100) was obtained for a unilateral loop, in most practical circuits, the bilateral correction factor is small, making (5.100) a good approximation of the loop gain even for a bilateral loop.

An important observation is that for a single loop where *both* the forward amplifier and the feedback network are unilateral, the return ratio determined for a dependent source in the forward direction,  $T$ , (e.g., given by (5.64)) and the unilateral loop gain,  $\mathbb{T}$ , (e.g., given by (5.92)) are the same. While there are many low frequency examples in which the amplifier is considered unilateral, in practice, they become bilateral at higher frequencies due to parasitic elements, such as capacitors. The generally bilateral loop will be discussed in section 5.4.2.

<sup>27</sup>This result for a unilateral loop gain is due to Rosenstark

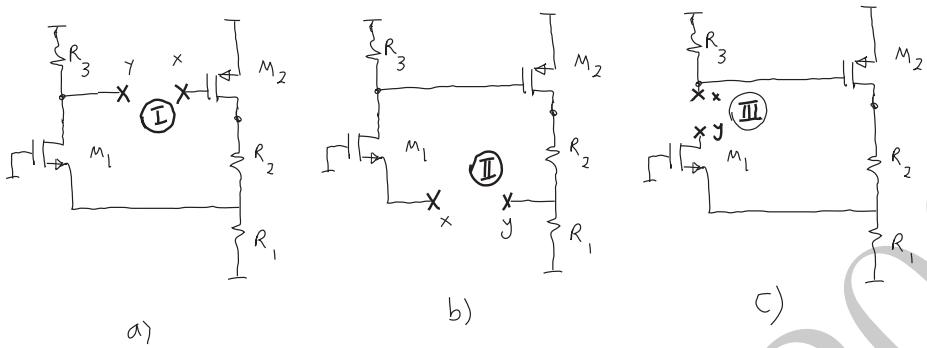


Figure 5.42: Breaking the loop of the series-shunt amplifier of Figure 5.4 at three different locations.

### Example 5.4.1 (MOS series-shunt feedback amplifier: Low Frequency Loop Gain)

Let us revisit the MOS series-shunt amplifier of Example 5.1.2 shown in Figure 5.4. In this example, we will calculate the low frequency forward loop gain of this amplifier using (5.100). We notice that no signal can flow in the reverse direction in the loop due to the low-frequency unilaterality of the transistors

Nulling the independent input source, we can break the loop at several different points. Let us start by breaking the loop at point I just before the gate of  $M_2$ , as shown in Figure 5.42a.

To calculate the (forward) voltage loop gain,  $\mathbb{T}_v$ , we apply a small-signal voltage source  $v_x$  to the gate of  $M_2$ , which results in a small-signal current  $g_{m2}v_x$  through its drain. This current goes through  $R_2$  and is pulled out of the parallel combination of  $R_1$  and the impedance seen looking into the source of  $M_1$ , namely,  $r_{m1}$ . Therefore, the voltage at the source of  $M_1$  is  $-g_{m2}(R_1 \parallel r_{m1})v_x$ . This voltage is then amplified by  $M_1$  which for this signal is in common-gate configuration, with a gain  $g_{m1}R_3$  from its source to its drain, which is  $v_y$ . Therefore,

$$\mathbb{T}_v \equiv -\frac{v_y}{v_x}|_{i_y=0} = g_{m1}g_{m2}R_3(R_1 \parallel r_{m1})$$

Next we need to calculate the (forward) current gain. This is done by applying a small-signal test current source,  $i_x$ , to the gate of  $M_2$ , and short circuiting the drain of  $M_2$  to ground. However, the low-frequency impedance seen looking into the gate of  $M_2$  is infinity. Therefore, the small-signal voltage at its gate is infinity and hence its drain current, the source voltage of  $M_2$ , and the drain current of  $M_2$  are all infinity as well, resulting in  $i_y$  being infinity. Thus,

$$\mathbb{T}_i \equiv -\frac{i_y}{i_x}|_{v_y=0} = \infty$$

Now combining the results using (5.100), we obtain:

$$\frac{1}{\mathbb{T}_I} = \frac{1}{\mathbb{T}_{v,I}} + \frac{1}{\mathbb{T}_{i,I}} = \frac{1}{\mathbb{T}_{v,I}}$$

hence

$$\mathbb{T}_I = \mathbb{T}_{v,I} = g_{m1}g_{m2}R_3(R_1 \parallel r_{m1})$$

where the subscript **I** represents the breakpoint of the loop, shown in 5.42a.

As we can see, breaking the loop at point **I** in Figure 5.42a results in one of the two forward loop gains (in this case,  $\mathbb{T}_i$ ) to become infinite. Although such a point may not exist in many practical circuits, identifying such points (or close approximations of), can simplify the calculations since we can determine the loop gain with a single calculation (e.g.,  $\mathbb{T}_v$  for point **I**).

Next we will see how the choice of the loop break-point results in different values for  $\mathbb{T}_v$  and  $\mathbb{T}_i$  while the final forward loop gain,  $\mathbb{T}$ , calculated in (5.100) is the same in end.

Now let us assume that we break the loop between the mid-point of the resistive feedback network and the source of  $M_1$  (point **II**), as illustrated in Figure 5.42b. First, we apply a small-signal test voltage,  $v_x$ , to the source of  $M_1$  and measured the returned voltage at the other end which is open circuited. The small-signal voltage at the drain of  $M_1$  is determined by the common-gate gain to be  $g_{m1}R_3$ . Therefore,  $M_2$  sinks a current  $g_{m2}g_{m1}R_3v_x$  out of the series combination of  $R_1$  and  $R_2$  which produces a voltage  $v_y = -g_{m2}g_{m1}R_1R_3v_x$ , across  $R_1$ . Hence,

$$\mathbb{T}_{v,II} \equiv -\frac{v_y}{v_x}|_{i_y=0} = g_{m2}g_{m1}R_1R_3$$

which is clearly different from  $\mathbb{T}_{v,I}$  calculated earlier.

Now to calculate  $\mathbb{T}_i$  we apply a test current source  $i_x$  to the source of  $M_1$  and short-circuit the other side (the common node between  $R_1$  and  $R_2$ ) to ground and measure the returned current,  $i_y$ . This time the drain voltage on  $M_1$  is  $i_xR_3$  and hence the drain current of  $M_2$  and thus the returned current is  $i_y = -g_{m2}R_3i_x$  resulting in

$$\mathbb{T}_{i,II} \equiv -\frac{i_y}{i_x}|_{v_y=0} = g_{m2}R_3$$

The total forward loop gain can be calculated using (5.100) to be

$$\frac{1}{\mathbb{T}_{II}} = \frac{1}{\mathbb{T}_{v,II}} + \frac{1}{\mathbb{T}_{i,II}} = \frac{1}{g_{m2}R_3} \cdot \left( \frac{1}{g_{m1}R_1} + 1 \right)$$

therefore

$$\mathbb{T}_{II} = g_{m1}g_{m2}R_3(R_1 \parallel r_{m1}) = \mathbb{T}_I$$

which is the same result obtained in the previous example breaking the loop at point **I**.

It is instructive to break the loop yet at another point: the drain of  $M_1$ , shown as point **III** in Figure 5.42c. Applying a voltage source,  $v_x$ , to  $R_3$  result in a finite small-signal voltage (proportional to  $v_x$ ) at the gate of  $M_2$  and hence the source of  $M_1$ . However, this time the drain of  $M_1$  is open circuited from a small-signal perspective and thus assuming  $r_{o1} \rightarrow \infty$ , it results in an infinite returned voltage,  $v_y$ , with a finite gate voltage. Therefore,

$$\mathbb{T}_{v,III} \equiv -\frac{v_y}{v_x}|_{i_y=0} = \infty$$

To calculate,  $\mathbb{T}_i$ , we apply a test current source  $i_x$  to  $R_3$  while the drain of  $M_1$  is shorted to a small-signal ground. This results in a gate voltage of  $i_x R_3$  for  $M_2$ , which in turn generates a drain current of  $g_{m2} R_3 i_x$  for  $M_2$ . This current is divided between  $R_1$  and  $r_{m1}$  with a factor of  $R_1/(R_1 + r_{m1})$ , which in turn becomes the drain current of  $M_1$  to produce  $i_y$ . Therefore, we have,

$$\mathbb{T}_{III} = \mathbb{T}_{i,III} \equiv -\frac{i_y}{i_x}|_{v_y=0} = g_{m2} R_3 \frac{R_1}{R_1 + r_{m1}} = g_{m1} g_{m2} R_3 (R_1 \parallel r_{m1})$$

that indicates

$$\mathbb{T}_I = \mathbb{T}_{II} = \mathbb{T}_{III}$$

In fact there are other points where the loop can be broken for which the reader can evaluate the same total forward loop gain.

The last example demonstrates that although different breakpoints for the loop do result in different individual values for the current and voltage forward loop gains, the overall *total* loop gain calculated using (5.100) will come out to be the same and is an *invariant of the loop*.

#### 5.4.2 Loop Gain of a Bilateral Loop

ADVANCED TOPIC

Online YouTube lecture:

[Bilateral loop gain, derivation, feedback](#)

It is important for the parameter defined as the “loop gain” in the general case to have certain properties. Loosely speaking we would like the “loop gain”, denoted by  $\mathbb{T}$  to be a generalization of the parameter  $Af$  in the idealized model of Figure 5.36a. From (5.2) it is clear that the closed-loop transfer function has a  $1 + Af$  term in the denominator, implying that we would like the transfer function to be inversely proportional to  $1 + \mathbb{T}$ . At the same time, we saw in (2.23) of Chapter 2 that the denominator of the transfer function is also inversely proportional to the determinant of the  $Y$  matrix, shown as  $\Delta$ . (This is why the poles of a circuit transfer function are the same as the roots of its determinant.) Therefore, a necessary condition for any parameter presented as “loop gain” is to satisfy,

$$(\mathbb{T} = -1) \Leftrightarrow (\Delta = 0) \quad (5.101)$$

In other word,  $1 + \mathbb{T}$  must have the same roots as  $\Delta$  to obtain the same set of transfer function poles. We will see in the next Chapter that this is an important necessary condition for the “loop gain” when it is used to determine the stability of feedback loops. However, there several ways to define  $\mathbb{T}$  to satisfying (5.101)<sup>28</sup> and thus  $\mathbb{T}$  cannot be uniquely defined just based on (5.101)<sup>28</sup>.

---

<sup>28</sup>For instance, the general definition of the return ratio given by (5.65) also satisfies this necessary condition.

For  $\mathbb{T}$  to be a generalization of  $Af$ , it is important for it to be invariant with where it is measured as long as we remain within the same loop<sup>29</sup>. This is not generally true for different parameters that satisfy (5.101). For instance, the return ratio defined in Section 5.2 satisfies (5.101) by the general definition of (5.65). However, as we saw at the beginning of Section 5.4 via (5.62) and (5.63) that it is *not* invariant of where it is measured even for a single bilateral loop. When it comes to the stability analysis of bilateral feedback loops, return ratio's lack of invariance immediately begs the question as to which one should be used to determine the loops stability<sup>30</sup>.

In addition to the above two traits, the general “loop gain” should also be easy to determine through a simple procedure for the feedback viewpoint to provide useful design insight and have merits over the brute force calculation using nodal equations or simulation of the transfer function. We now embark on a definition of loop gain that satisfies these three requirements for the generally bilateral case and reduces to the earlier results for unilateral circuits. To be able to do so, we first define the *reverse* voltage and current loop gains in a bilateral loop.

### Reverse Voltage and Current Loop Gain

In practice, there is always a certain amount of reverse transmission in any block. As a results, in addition to the forward (clockwise in Figures 5.36a) loop gain,  $\mathbb{T}_f$ , discussed in the previous subsection<sup>31</sup>, there is also a *reverse* loop gain,  $\mathbb{T}_r$  with the signal traveling in the opposite direction in the loop (i.e., going backward in the amplifier and in reverse through the feedback network).

Similar to the forward loop gain, we can define the reverse current and voltage loop gains by separately applying a test voltage,  $v_y$ , and current source,  $i_y$ , in the reverse direction (counterclockwise in Figures 5.36a), opening and shorting the other end of the loop, respectively, as shown in Figure 5.43. Measuring the reverse returned voltage,  $v_x$ , and current,  $i_x$ , we can define the reverse voltage and current loop gains as

$$\mathbb{T}_{v,r} \equiv -\frac{v_x}{v_y}|_{i_x=0} \quad (5.102a)$$

$$\mathbb{T}_{i,r} \equiv -\frac{i_x}{i_y}|_{v_x=0} \quad (5.102b)$$

Note that in this case, the  $i_y$  and  $v_y$  are the independent excitation sources and the measured returned quantities are  $i_x$  and  $v_x$ , respectively. It is very important to note that the reverse loop gains are *not* simply the inverse of the

---

<sup>29</sup>Sometimes it might not be obvious whether we are still within the same major loop, if we are dealing with parasitic elements in transistors. We will talk more about this in subsection ??.

<sup>30</sup>If the parameter is not invariant but satisfies (5.101) it can result in different zeros although it produces the same poles.

<sup>31</sup>In this subsection, we will differentiate between the forward and reverse gains by using the appropriate subscripts,  $f$  and  $r$ . If neither of these subscripts are presents it should be assumed we are talking about the forward loop gain unless explicitly stated otherwise.

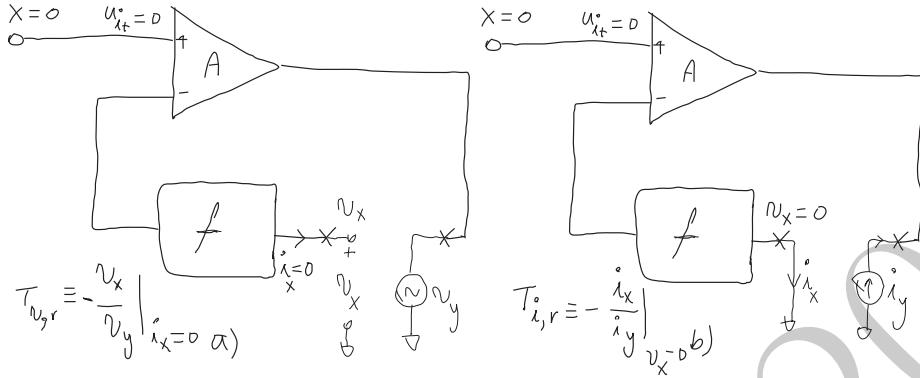


Figure 5.43: Breaking the loop for calculation of the reverse a) voltage loop gain, and b) current loop gain.

forward loop gains defined in (5.93) and (5.94) because the excitation is applied at a different place and different nodes are opened or shorted.

While in many cases, the amplifier's reverse gain is small compared to its forward gain at low frequencies, it is important to know how to take this effect into account in the cases where it does matter. To see how this is treated in general, we go back to the general two-port model of Figure 5.30, which allows both the amplifier and the feedback networks to have reverse transmissions ( $G_1$  and  $G_4$ ). In this case, both forward and reverse loop gains would be non-zero. Next we will derive an expression to calculate the general loop gain in terms of forward and reverse, voltage and current loop gains.

The transfer function of the general bilateral network shown in Figure 5.30, can be calculated via nodal analysis<sup>32</sup>, namely,

$$Y \cdot V = \begin{pmatrix} Y_i & G_i \\ G_o & Y_o \end{pmatrix} \cdot \begin{pmatrix} v_i \\ v_o \end{pmatrix} = \begin{pmatrix} i_{in} \\ 0 \end{pmatrix} \quad (5.103)$$

Hence, the determinant of the circuit is given by<sup>33</sup>

$$\Delta = Y_i Y_o - G_i G_o = (Y_1 + Y_3)(Y_2 + Y_4) - (G_1 + G_3)(G_2 + G_4) \quad (5.104)$$

Next we will demonstrate that the forward loop expression for the unilateral gain given in (5.100) does not exactly satisfy (5.101) in its present form. To do so we will need to determine the forward voltage and current loop gains for

<sup>32</sup>Writing KCL at the input and output nodes, i.e.,

$$\begin{aligned} i_{in} &= (Y_1 + Y_3)v_i + (G_1 + G_3)v_o = Y_i v_i + G_i v_o \\ 0 &= (G_2 + G_4)v_i + (Y_2 + Y_4)v_o = G_o v_i + Y_o v_o \end{aligned}$$

<sup>33</sup>Remembering that  $Y_i = Y_1 + Y_3$ ,  $Y_o = Y_2 + Y_4$ ,  $G_i = G_1 + G_3$ , and  $G_o = G_2 + G_4$ .

## ▼ Derivation ▼

the bilateral loop of Figure 5.30. We can break the loop at any point. Let us assume that the loop is broken at the output.

To calculate  $\mathbb{T}_{v,f}$ , similar to the unilateral case in the previous subsection, we can write the KCL at the input and output nodes to obtain

$$G_2v_i + Y_2v_y = 0 \quad (5.105a)$$

$$Y_i v_i + G_1 v_y + G_3 v_x = 0 \quad (5.105b)$$

which leads to

$$\mathbb{T}_{v,f} \equiv -\frac{v_y}{v_x}|_{i_y=0} = -\frac{G_2G_3}{Y_iY_2 - G_1G_2} \quad (5.106)$$

The forward current gain is calculated by applying the KCL at the three nodes of the circuit leading to

$$Y_4v_1 + G_4v_i = i_x \quad (5.107a)$$

$$G_2v_i = -i_y \quad (5.107b)$$

$$Y_i v_i + G_3 v_1 = 0 \quad (5.107c)$$

where there is an extra,  $G_4v_i$ , term in the first equation due to the reverse path in the feedback path. Therefore, we have,

$$\mathbb{T}_{i,f} \equiv -\frac{i_y}{i_x}|_{v_y=0} = -\frac{G_2G_3}{Y_iY_4 - G_3G_4} \quad (5.108)$$

Now let us see whether the forward loop gain expression for unilateral loop in (5.100) satisfies (5.101).

For the bilateral loop, using (5.106) and (5.108), we can calculate the sum of the reciprocals of the forward voltage and current loop gains,

$$\frac{1}{\mathbb{T}_{v,f}} + \frac{1}{\mathbb{T}_{i,f}} = -\frac{Y_iY_o - G_1G_2 - G_3G_4}{G_2G_3} = -\frac{\Delta + G_2G_3 + G_1G_4}{G_2G_3} \stackrel{?}{=} \frac{1}{\mathbb{T}_f} \quad (5.109)$$

It is not clear whether the above quantity is in fact the actual forward loop-gain, as indicated by the question mark in the last equality (which we will shortly see is not true). It is easy to see that  $\mathbb{T}_f = -1$  indicates  $\Delta = -G_1G_4$ , which results in  $\Delta = 0$  *only* for a unilateral loop (when at least one of  $G_1$  and  $G_4$  is zero). Although, this verifies the validity of (5.100) for the unilateral case, it also shows that it fails in the bilateral case.

However, we notice that if we modify (5.109) so that the term  $G_1G_4$  is not present in it anymore, the resultant expression will indeed satisfy (5.101) which indicates that the forward loop gain expression could be expressed as

$$\mathbb{T}_f = -\frac{G_2G_3}{\Delta + G_2G_3} \quad (5.110)$$

Now we can modify (5.100) for the bilateral loop by adding a correction factor the sum of the reciprocals of the forward voltage and current gains, i.e.,

$$\frac{1}{\mathbb{T}_f} = -\frac{\Delta + G_2G_3}{G_2G_3} = \frac{1}{\mathbb{T}_{v,f}} + \frac{1}{\mathbb{T}_{i,f}} + \frac{G_1G_4}{G_2G_3} \quad (5.111)$$

Now we can determine the last term by measuring the *reverse* voltage or current loop gains. The *reverse* voltage loop gain is calculated by applying KCL to obtain

$$G_4v_i + Y_4v_x = 0 \quad (5.112a)$$

$$Y_i v_i + G_1 v_y + G_3 v_x = 0 \quad (5.112b)$$

leading to

$$\mathbb{T}_{v,r} \equiv -\frac{v_x}{v_y}|_{i_x=0} = -\frac{G_1 G_4}{Y_i Y_4 - G_3 G_4} \quad (5.113)$$

Finally, the *reverse* current gain is computed using the following nodal equations,

$$G_4v_i = -i_x \quad (5.114a)$$

$$G_2v_i + Y_2v_o = i_y \quad (5.114b)$$

$$Y_i v_i + G_1 v_o = 0 \quad (5.114c)$$

resulting in

$$\mathbb{T}_{i,r} \equiv -\frac{i_x}{i_y}|_{v_x=0} = -\frac{G_1 G_4}{Y_i Y_2 - G_1 G_2} \quad (5.115)$$

An interesting observation at this point is that

$$\boxed{\mathbb{T}_{v,f} \cdot \mathbb{T}_{v,r} = \mathbb{T}_{i,f} \cdot \mathbb{T}_{i,r}} \quad (5.116)$$

which indicates that if we know any three of the forward and reverse voltage and current loop gains ( $\mathbb{T}_{v,f}$ ,  $\mathbb{T}_{v,r}$ ,  $\mathbb{T}_{i,f}$ , and  $\mathbb{T}_{i,r}$ ), we can calculate the fourth one. This is a useful intermediate result.

However, comparing (5.106) and (5.108) (or equivalently (5.113) and (5.115) because of (5.116) equality), we can rewrite the bilateral correction factor in (5.111) as:

$$\frac{G_1 G_4}{G_2 G_3} = \frac{\mathbb{T}_{i,r}}{\mathbb{T}_{v,f}} = \frac{\mathbb{T}_{v,r}}{\mathbb{T}_{i,f}} \quad (5.117)$$

which can be used to calculate the total forward loop gain.

Therefore, the total effective forward loop gain for the general case of bilateral feedback and amplifier is given by:

$$\boxed{\frac{1}{\mathbb{T}_f} = \frac{1}{\mathbb{T}_{v,f}} + \frac{1}{\mathbb{T}_{i,f}} + \frac{\mathbb{T}_{i,r}}{\mathbb{T}_{v,f}} = \frac{1}{\mathbb{T}_{v,f}} + \frac{1}{\mathbb{T}_{i,r}} + \frac{\mathbb{T}_{v,r}}{\mathbb{T}_{i,f}}} \quad (5.118)$$

which reduces to (5.100) if either of the reverse loop gains ( $\mathbb{T}_{v,r}$  or  $\mathbb{T}_{i,r}$ ) are zero.

similarly we can determine the effective reverse loop gain (though it is typically less significant). Total effective reverse loop gain:

$$\frac{1}{\mathbb{T}_r} = \frac{1}{\mathbb{T}_{v,r}} + \frac{1}{\mathbb{T}_{i,r}} + \frac{\mathbb{T}_{i,f}}{\mathbb{T}_{v,r}} = \frac{1}{\mathbb{T}_{v,r}} + \frac{1}{\mathbb{T}_{i,r}} + \frac{\mathbb{T}_{v,f}}{\mathbb{T}_{i,r}} \quad (5.119)$$

## ▼ Result ▼

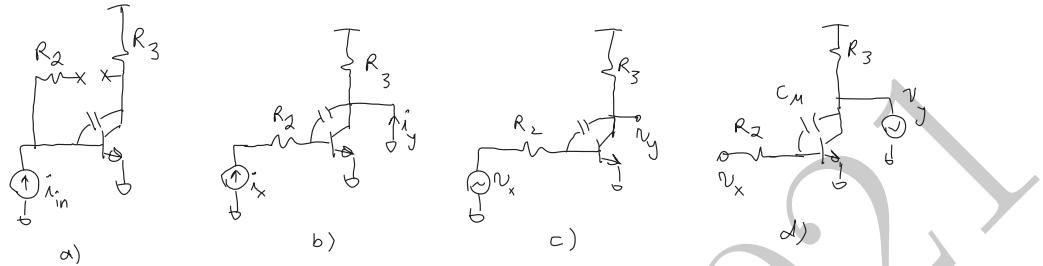


Figure 5.44: a) Breaking the feedback loop for the shunt-shunt amplifier of Figure 5.5a. b) Calculation of the forward current loop gain,  $\mathbb{T}_{i,f}$ , c) the forward voltage loop gain,  $\mathbb{T}_{v,f}$ , and d) the reverse voltage loop gain,  $\mathbb{T}_{v,r}$ .

**Example 5.4.2** Let us revisit the BJT shunt-shunt amplifier of Figure 5.5a. This time let us take the  $C_\mu$  of the transistor into account. If we break the loop between the output and the  $R_2$  as shown in Figure 5.44a, we see that at high frequencies, we have non-zero reverse transmission due to  $C_\mu$  and hence the loop is bilateral. Let us first calculate the forward voltage and current loop gains.

To calculate the forward current loop-gain,  $\mathbb{T}_{i,f}$ , we apply a small-signal test current source,  $i_x$ , to  $R_2$  and short-circuit the collector to ground (from a small-signal perspective) to determine  $i_y$ , as shown in Figure 5.44b. We could use the nodal analysis or the method of time constants to determine the transfer function, using the method of time constants we have,

$$\tau_{i,f}^0 = r_\pi C_\mu \quad (5.120)$$

$$\mathbb{T}_{i,f}^0 = \beta \quad (5.121)$$

$$\mathbb{T}_{i,f}^\mu = -1 \quad (5.122)$$

where  $\tau_{i,f}^0$  is easily determined noting that  $C_\mu$  is shorted to ground on the other side and is hence in parallel with  $r_\pi$ . It is also easy to see that for  $C_\mu = 0$ , current  $i_x$  is simply amplified by  $\beta$  of the transistor. To Calculate  $\mathbb{T}_{i,f}^\mu$ , we set  $C_\mu$  to its infinite value, i.e., short circuit. In this case, the  $i_x$  directly appears as  $i_y$  and hence because of the minus sign in the definition of loop gain, we obtain  $\mathbb{T}_{i,f}^\mu = -1$  to arrive at the forward current loop gain as a function of frequency,

$$\mathbb{T}_{i,f}(s) = \frac{\mathbb{T}_{i,f}^0 + \mathbb{T}_{i,f}^\mu \tau_{i,f}^0 s}{1 + \tau_{i,f}^0 s} = \mathbb{T}_{i,f}^0 \cdot \frac{1 - \tau_z s}{1 + \tau_{i,f}^0 s} \quad (5.123)$$

where  $\tau_z = r_m C_\mu$ .

Now we can determine the forward voltage loop gain,  $\mathbb{T}_{v,f}$  by applying a voltage source to  $R_2$  and leaving the other end open circuited, as in Figure

5.44c. This time we have,

$$\begin{aligned}\tau_{v,f}^0 &= [(r_\pi \parallel R_2) + R_3 + g_m R_3 (r_\pi \parallel R_2)] C_\mu \\ &= \frac{r_\pi C_\mu}{R_2 + r_\pi} \cdot (R_2 + R_3 + \frac{R_2 R_3}{\alpha r_m})\end{aligned}\quad (5.124)$$

$$\mathbb{T}_{v,f}^0 = \frac{r_\pi}{R_2 + r_\pi} \cdot g_m R_3 = \frac{\beta R_3}{R_2 + r_\pi} \quad (5.125)$$

$$\mathbb{T}_{v,f}^\mu = -\frac{\alpha r_m \parallel R_3}{R_2 + \alpha r_m \parallel R_3} = -\frac{R_3}{R_2 + R_3 + \frac{R_2 R_3}{\alpha r_m}} \quad (5.126)$$

where  $\tau_{v,f}^0$  is determined by applying (3.63) of Chapter (3), as this time the resistance on the left of  $C_\mu$  is  $R_2 \parallel r_\pi$  and  $R_3$  on its righthand side.  $\mathbb{T}_{v,f}^0$  is simply the low-frequency gain of common-emitter, and  $\mathbb{T}_{v,f}^\mu$  is determined by the voltage divider between  $\alpha r_m \parallel R_3$  and  $R_2$  (remember that the resistance of a diode connected BJT is  $\alpha r_m$ ). Based on the above, we have

$$\mathbb{T}_{v,f}(s) = \frac{\mathbb{T}_{v,f}^0 + \mathbb{T}_{v,f}^\mu \tau_{v,f}^0 s}{1 + \tau_{v,f}^0 s} = \mathbb{T}_{v,f}^0 \cdot \frac{1 - \tau_z s}{1 + \tau_{v,f}^0 s} \quad (5.127)$$

where  $\tau_z = r_m C_\mu$  similar to  $\mathbb{T}_{v,f}(s)$ .

We know that we have a bilateral loop for the above loop gain, so we need to determine at least one of the reverse loop gains (the other one is redundant because of (5.116)) to be able to determine the overall loop gain. We calculate  $\mathbb{T}_{v,r}$  here. To do so, we apply a test voltage,  $v_y$ , to the collector and measure the voltage at the open-circuited end of  $R_2$ , as shown in Figure 5.44d. This can be easily done, by using the GTC, i.e.,

$$\tau_{v,r}^0 = r_\pi C_\mu = \tau_{i,f}^0 \quad (5.128)$$

$$\mathbb{T}_{v,r}^0 = 0 \quad (5.129)$$

$$\mathbb{T}_{v,r}^\mu = -1 \quad (5.130)$$

where the reverse voltage loop gain in the absence of  $C_\mu$  is zero. Based on this we have

$$\mathbb{T}_{v,r}(s) = \frac{\mathbb{T}_{v,r}^0 + \mathbb{T}_{v,r}^\mu \tau_{v,r}^0 s}{1 + \tau_{v,r}^0 s} = \frac{-\tau_{v,r}^0 s}{1 + \tau_{v,r}^0 s} = \frac{-\tau_{i,f}^0 s}{1 + \tau_{i,f}^0 s} \quad (5.131)$$

which verifies that at low-frequencies the loop is unilateral ( $\mathbb{T}_{v,r}(0) = 0$ ).

Now let us combine the results of (5.123), (5.127), and (5.131), using (5.118), we obtain,

$$\begin{aligned}\frac{1}{\mathbb{T}_f(s)} &= \frac{1}{\mathbb{T}_{v,f}(s)} + \frac{1}{\mathbb{T}_{i,f}(s)} + \frac{\mathbb{T}_{v,r}(s)}{\mathbb{T}_{i,f}(s)} = \frac{1}{1 - \tau_z s} \left[ \frac{1 + \tau_{v,f}^0 s}{\mathbb{T}_{v,f}^0} + \frac{1}{\mathbb{T}_{i,f}^0} \right] \\ &= \left( \frac{1}{\mathbb{T}_{v,f}^0} + \frac{1}{\mathbb{T}_{i,f}^0} \right) \cdot \frac{1 + \frac{\tau_{v,f}^0 \mathbb{T}_{i,f}^0}{\mathbb{T}_{v,f}^0 + \mathbb{T}_{i,f}^0} s}{1 - \tau_z s} = \frac{1}{\mathbb{T}_f(0)} \cdot \frac{1 + \tau_1 s}{1 - \tau_z s}\end{aligned}\quad (5.132)$$

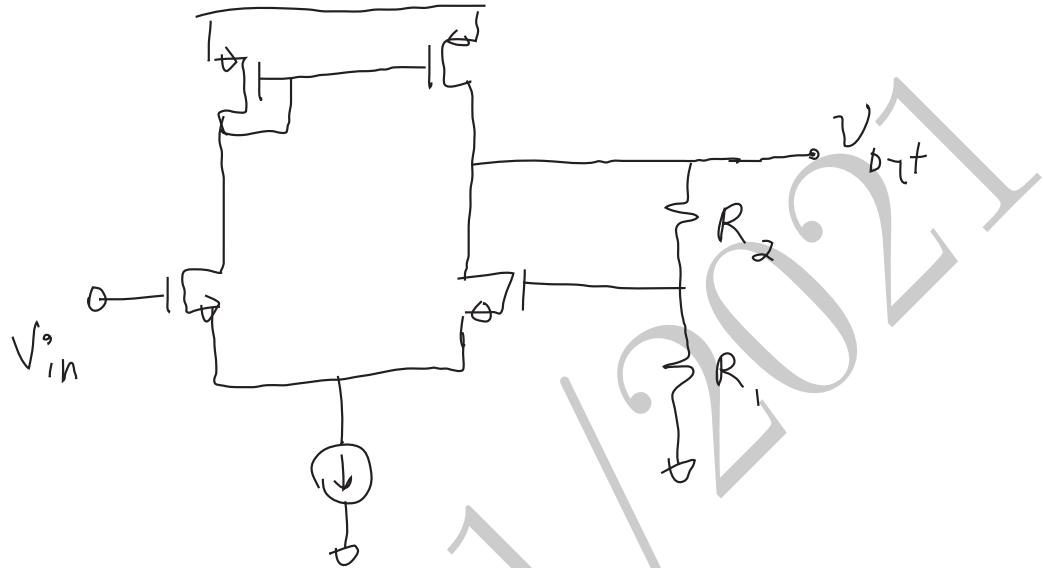


Figure 5.45: A single-stage MOS op-amp in series-shunt feedback configuration.

where  $\tau_{v,r}^0 = \tau_{i,f}^0$  was used to arrive at the third equality. Note that the low frequency forward loop gain is simply given by the unilateral expression, i.e., (5.100) since the loop is unilateral at low frequencies. This leads to

$$T_f(s) = \frac{\beta R_3}{r_\pi + R_2 + R_3} \cdot \frac{1 - r_m C_\mu s}{1 + [r_\pi \parallel (R_2 + R_3) + \frac{(1+\beta)R_2 R_3}{R_2 + R_3 + r_\pi}] C_\mu s} \quad (5.133)$$

which for  $\beta \rightarrow \infty$  (e.g., MOSFET), reduces to

$$T_f(s) = g_m R_3 \cdot \frac{1 - r_m C_\mu s}{1 + (R_2 + R_3 + g_m R_2 R_3) C_\mu s} \quad (5.134)$$

Figure 5.45.

Figure 5.46.

Figure 5.47.

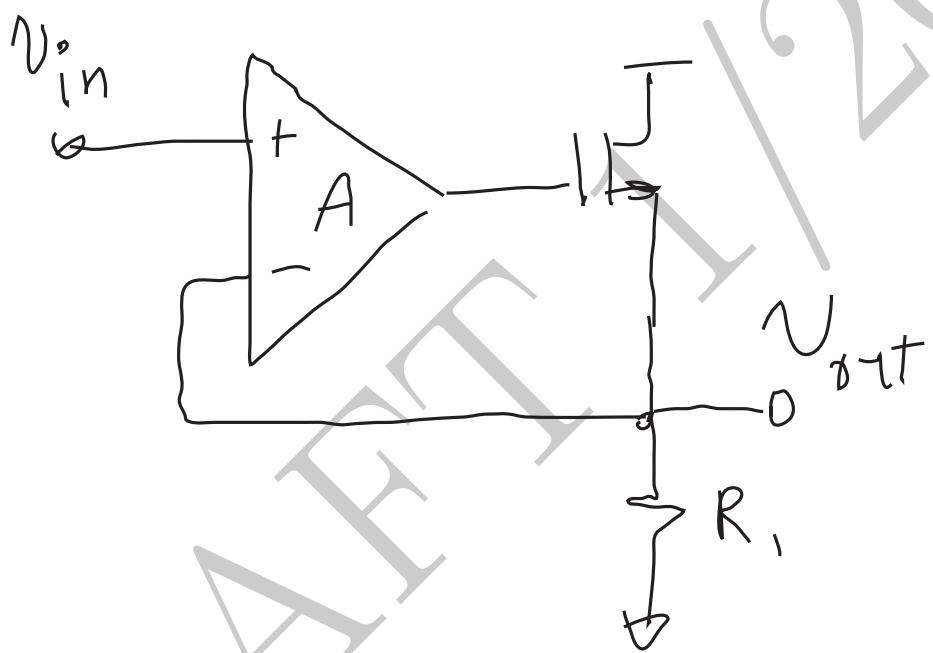


Figure 5.46: A MOSFET series-shunt feedback circuit, super buffer.

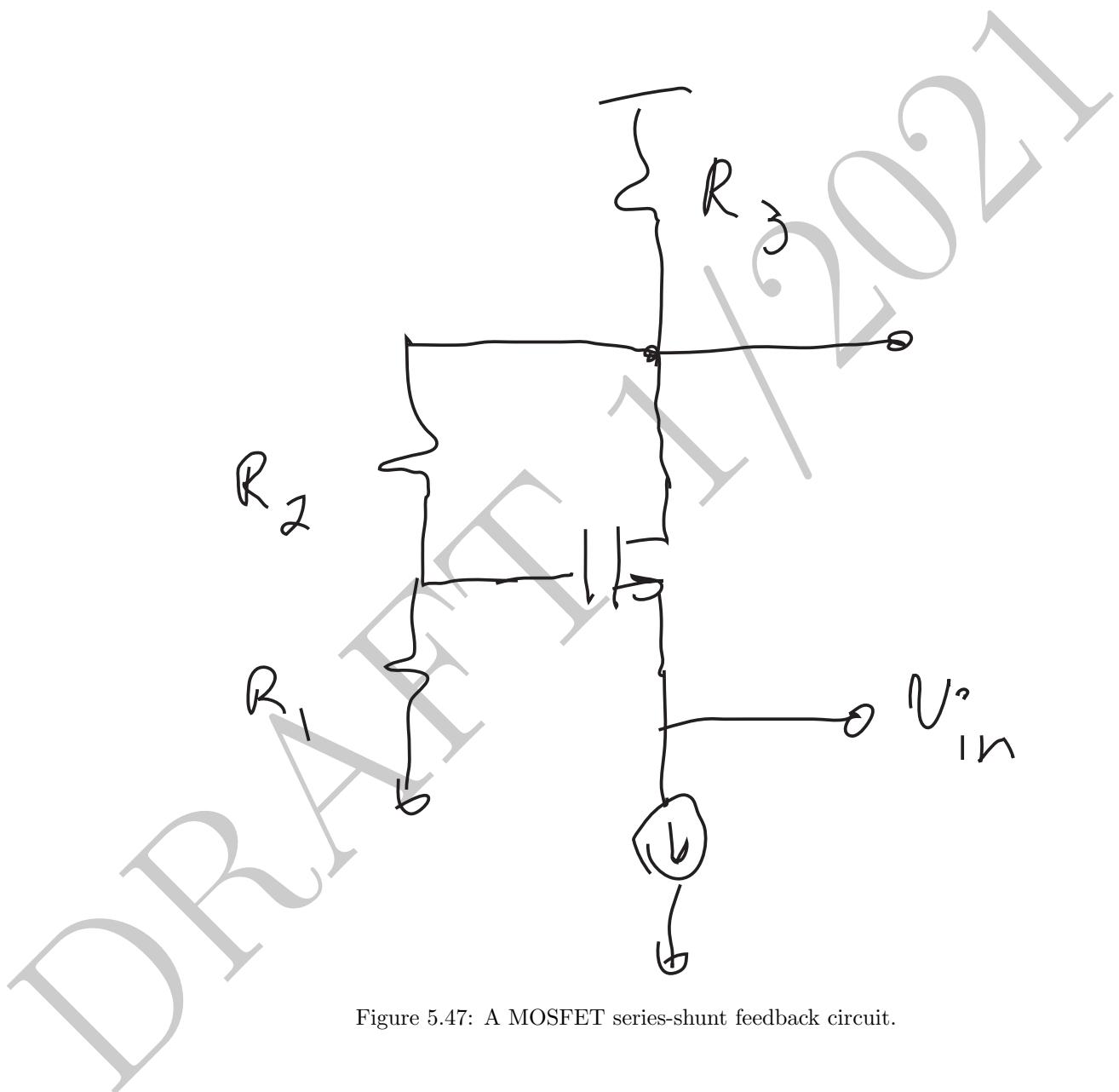


Figure 5.47: A MOSFET series-shunt feedback circuit.

## Chapter 6

# Stability and Compensation

We saw in Chapter 5 that feedback offered several benefits including accurate control of the closed-loop gain, desensitization, and linearization of the forward path amplifier. The feedback could be quantified with different metrics two of which were the return ratio, shown as  $T$ , and the loop gain, designated as  $\bar{T}$ . We also noticed that the amount of improvement due to feedback was directly related to the amount of the signal that was fed back. Thus, it appears that the larger the loop gain, the more plentiful the benefits of feedback. However, excessive feedback in the circuit can result in some behavior that may not be initially easily visible. Some of these behavior can be undesirable for some applications (e.g., amplifier instability), while they can provide useful new behavior for others (e.g., oscillators). We will demonstrate some of these behavior in an example and then discuss the general treatment.

Online YouTube lecture:

[Instability and conditional stability, circuit examples.](#)

**Example 6.0.1 (Forward Amplifier with Multiple Gain Stages)** Let us start with the trans-impedance amplifier of Figure 6.1, where three common-source amplifying stages are cascade to form a high gain forward amplifier. A buffer stage is introduced at the output to isolate the loading effect of different feedback networks. This amplifier could be used in an optical front-end to convert the output current of a reversed biased photo diode to an output voltage (FIXXX). For the time being let us assume that the dc bias has been established and simply look at the ac circuit. In the absence of feedback, the amplifier has a low-frequency forward trans-impedance of

$$a_0 \equiv Z(0) \equiv \frac{v_{out}}{i_{in}} = -g_{m1}R_1 \cdot g_{m2}R_2 \cdot g_{m3}R_3 \cdot R_S \quad (6.1)$$

For the time being let us only consider the input source capacitance,  $C_D$  (e.g., that of the photodiode) and  $C_{gs}$ 's of  $M_1$  through  $M_3$ . This results in three independent capacitors:  $C_{in} = C_D + C_{gs1}$ ,  $C_1 = C_{gs2}$ , and  $C_2 = C_{gs3}$  in

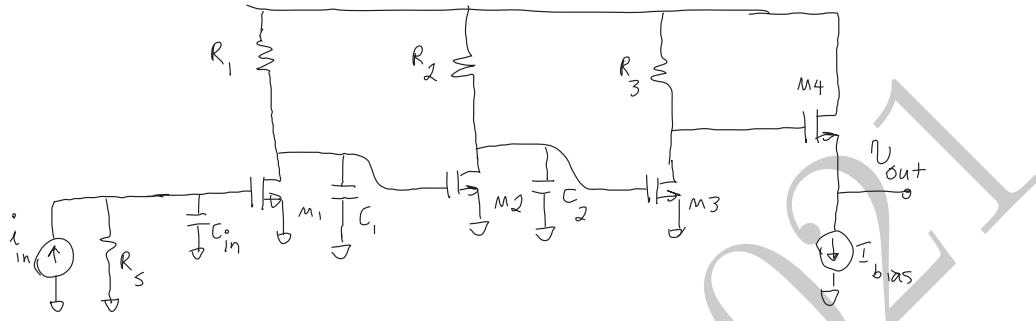


Figure 6.1: A three stage amplifier with output buffer.

*Figure 6.1.* In this example, we ignore  $C_{gd}$ 's of  $M_1$ ,  $M_2$ , and  $M_3$ , and the capacitors of the source-follower,  $M_4$ . It can be seen that under these simplifying assumptions and in the absence of a feedback, the three time constants associated with  $C_{in}$ ,  $C_1$ , and  $C_2$  are uncoupled, since shorting or opening of any of the capacitors will not change the resistance seen by the other ones<sup>1</sup>. Also, we note that short-circuiting none of these three capacitors results in a non-zero low-frequency transfer function, therefore there are no zeros when we only consider these three capacitors. Thus, the transfer function (trans-impedance) will simply be:

$$Z(s) \equiv \frac{v_{out}}{i_{in}} = \frac{a_0}{(1 + \tau_{in}s)(1 + \tau_1s)(1 + \tau_2s)} \quad (6.2)$$

where

$$\tau_{in} = R_S C_{in}$$

$$\tau_1 = R_1 C_1$$

$$\tau_2 = R_2 C_2$$

#### ♦ Numerical Example ♦

corresponding to three real LHP poles.

Let us assume  $a g_m = 2mS$  and  $C_{gs} = 500fF$  for all the transistors and load resistors ( $R_1$ ,  $R_2$ , and  $R_3$ ) of  $5k\Omega$ , and a source resistance and capacitance of  $R_S = 1k\Omega$  and  $C_D = 1pF$ , respectively. We have  $a_0 = 1M\Omega(120dB\Omega)$ . Also, we can calculate the time constants to be:  $\tau_{in} = 1.5ns$ ,  $\tau_1 = 2.5ns$ , and  $\tau_2 = 2.5ns$ . These correspond to one LHP pole at  $106MHz$  and two overlapping real poles at  $64MHz$ . The simulated magnitude of the frequency-domain transfer function and the step response of the amplifier are plotted in Figure 6.2, where we have a monotonically decreasing amplitude and phase responses, as well as a clean step response with no ringing.

We observe a  $3dB$  bandwidth of  $36.6MHz$ . ZVT's produce a conservative estimate of  $24.5MHz$ .

Now, we introduce a feedback resistor,  $R_F$  from the output back to the gate of  $M_1$  to create a shunt-shunt feedback configuration, as depicted in Figure 6.3.

<sup>1</sup>For detailed discussions on this see Chapter 4.

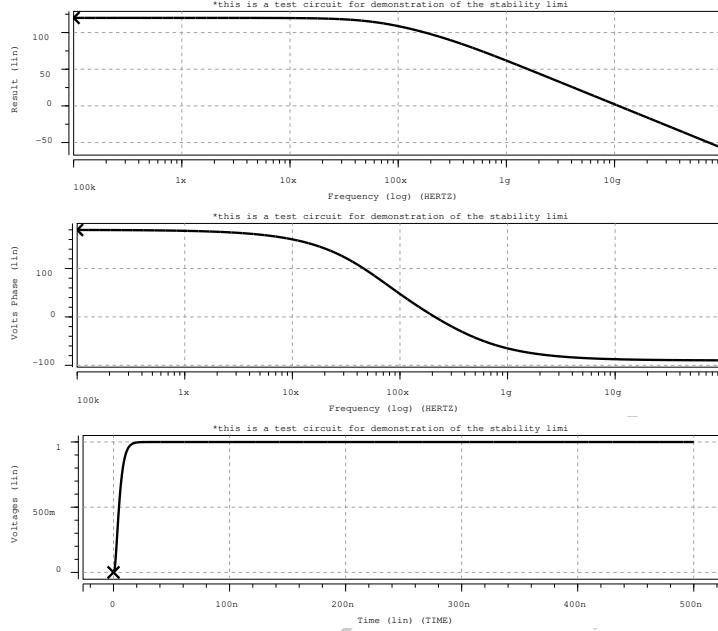


Figure 6.2: The magnitude and phase of the loop frequency-domain transfer function and the step response of the three-stage amplifier of Figure 6.1, in the absence of  $C_{ga}$ 's and capacitors of  $M_4$ .

We now reevaluate the transfer function for different values of  $R_F$ . It is easy to see that large  $R_F$  corresponds to small feedback since for  $R_F \rightarrow \infty$ , the stage reduces to the amplifier with no feedback. By reducing  $R_F$  we increase the feedback applied (and hence the loop gain).

First, let us evaluate the behavior with  $R_F = 500k\Omega$ . The simulated transfer function and time-domain step response are shown in Figure 6.4. As can be seen the feedback lowers  $a_0$  to  $110dB\Omega$ . We also note about  $4dB$  of peaking in the amplitude response around  $75MHz$ . In the time domain step response, we also notice some overshoot and a little bit of ringing.

Next let us increase the feedback by lowering  $R_F$  to  $150k\Omega$ . The frequency and time domain responses are shown in Figure 6.5, where we see that  $a_0$  is now reduced by the feedback to  $102dB\Omega$ . However, we also notice  $19dB$  of peaking in frequency domain! This corresponds to a very under-damped response as evident from the large ringing in the time domain step response.

Now if we increase the feedback further by lowering  $R_F$  to  $116k\Omega$ , we notice in Figure 6.6 that the  $a_0$  goes to  $100dB\Omega$  but we have see about  $53dB$  of peaking with a peak gain of  $153dB\Omega$  at  $133MHz$ ! The step response shows a very slowly decaying oscillation. It is clear that this highly-underdamped behavior presents

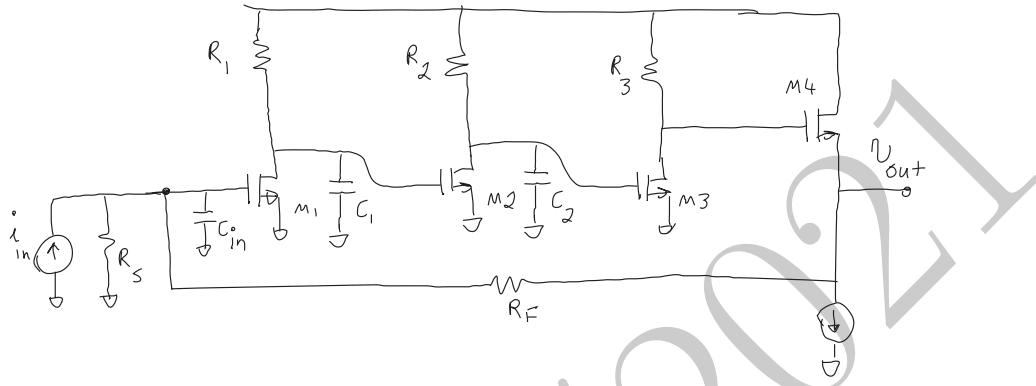


Figure 6.3: The three stage amplifier stage of Figure 6.1 under shunt-shunt resistive feedback.

a real challenge to an amplifier and effectively renders it useless as such.

At this point one probably wonders what happens if we increase the feedback by lowering  $R_F$  even further. Let us look at the frequency and time domain behavior of this circuit for a slightly lower  $R_F$  of  $112k\Omega$  (yet more feedback), as illustrated in Figure 6.7. The Bode plot does not seem to indicate anything too exciting. In fact, it seems to indicate that the small-signal transfer function has a smaller peaking of  $38dB$  compared to  $53dB$  of peaking in the case of  $R_F = 116k$ . However, a glance at the step response (plotted on a different time scale from the previous ones), reveals a totally different story. More specifically, we see that the amplitude of the oscillatory step response of the system grows (at least initially) with time as opposed to decaying with it. So this circuit will likely become an oscillator once the nonlinearity of the circuit elements kick in and limit the amplitude (more on the effect of nonlinearity later in XXX).

As an interesting special case, we try to find a value for  $R_F$  for which the ringing does neither decays, nor grows and thus stays constant. Let us call it the critical feedback resistor,  $R_{F,crit}$ . In our numerical example, with all  $C_{gd}$ 's and all the capacitors of the output buffer being zero, we find it to be  $R_{F,crit} \approx 115.4k\Omega$ . For this value, the step response shows a constant ringing at  $133MHz$ .

Now let us evaluate the loop gain under this critical value of  $R_F$ . The loop can be broken at many different places, as we discussed in Section ???. Considering that at this stage we have decided to ignore the parasitic capacitors of  $M_4$ , its gate presents a convenient place to break the loop since  $T_i \rightarrow \infty$ . Figure 6.8 illustrates how we can break the loop at the gate of  $M_4$  and apply a sinusoidal input voltage and look at the returned voltage to evaluate the loop gain (according to (5.100), we have  $T = T_v$  since  $T_i \rightarrow \infty$ ).

Figure 6.9 shows the input ( $v_x$ ) and the output ( $v_y$ ) voltages of the broken loop including the initial transient passes, for an input sinusoid of  $133MHz$ . As

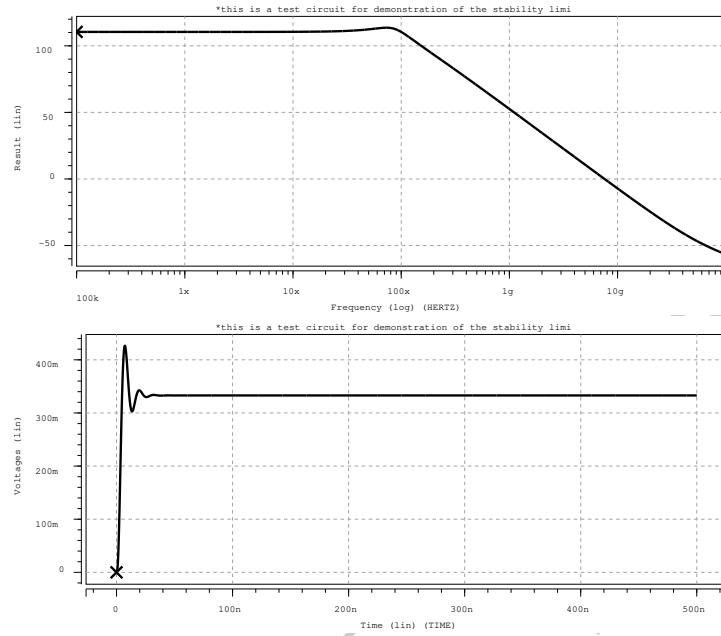


Figure 6.4: The closed-loop transfer function and the step response of the three-stage amplifier of Figure 6.3, with  $R_F = 500\text{k}\Omega$ .

can be seen in this case for  $R_F = R_{F,crit}$ , the input and output sinusoids have the same frequency and amplitude and have a time shift that is an integer multiple of the period (for a single frequency sine wave this would be a phase difference that is an integer multiple of  $360^\circ$  or  $2\pi$ ). Thus the substitution theorem implies that the voltage source  $v_x$  can be replaced with the drain of  $M_3$  and the circuit can maintain this waveform. In other words, this is a steady-state solution for the circuit which can be maintained. Therefore, if the circuit has a loop gain that is exactly  $-1$  (magnitude of 1 and phase of  $180^\circ$ ) at a given frequency<sup>2</sup>, the circuit can maintain steady-state oscillations at that frequency. However, one should be very careful that does not generally apply for cases where the loop gain has a magnitude smaller or greater than unity with a phase of  $180^\circ$ .

Later we will see examples of linear systems where the returned sinusoid is in phase and greater in amplitude than the injected signal, where the closed loop system does not show growing or sustained oscillatory behavior. In fact in the early days of electronic circuits, people noticed feedback systems that oscillated for given amount of feedback, but became stable with increasing loop gain! This was the reason for development of the more general Nyquist stability criterion

<sup>2</sup>The negative sign is to compensate the minus sign in the definition of the loop gain, i.e., 5.91.

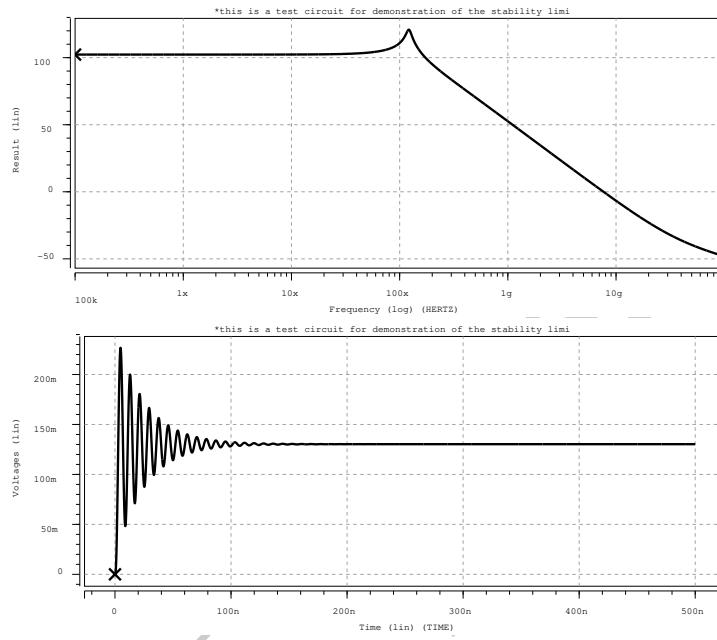


Figure 6.5: The closed-loop transfer function and the step response of the three-stage amplifier of Figure 6.3, with  $R_F = 150k\Omega$ .

*discussed in section 6.1.*

In the case of the circuit in Figure 6.3, however, we noticed what if we had chosen a smaller  $R_F$ , resulting in a larger feedback, we would observe a growing exponential behavior. We saw earlier that for instance for  $R_F = 112k\Omega$ , the oscillatory output of the circuit grows exponentially. For a broken loop, if we apply a sinusoidal input,  $v_x$ , at 133MHz, the output,  $v_y$  comes back with a total phase shift of  $2\pi$  (essentially in phase) and an amplitude that is 1.11 times greater than the input. In this example, it seems that a greater returned signal can indeed result in growing exponential. We will see examples later where the returned signal is in phase and has a greater amplitude, but does not result in an oscillatory output. We will discuss the conditions leading to this and the criteria to determine if and when this happens later in Section XXX.

Let us evaluate the loop gain. As discussed in Chapter 5, the loop gain can be determined by breaking the loop at the gate of  $M_4$  since it makes  $T_i$  go to infinity, so we using 5.100 we know the loop gain to be the voltage loop gain for that point. In this case, the low frequency loop gain is essentially the voltage gain of the three common-source stages times the voltage division ratio from  $v_x$

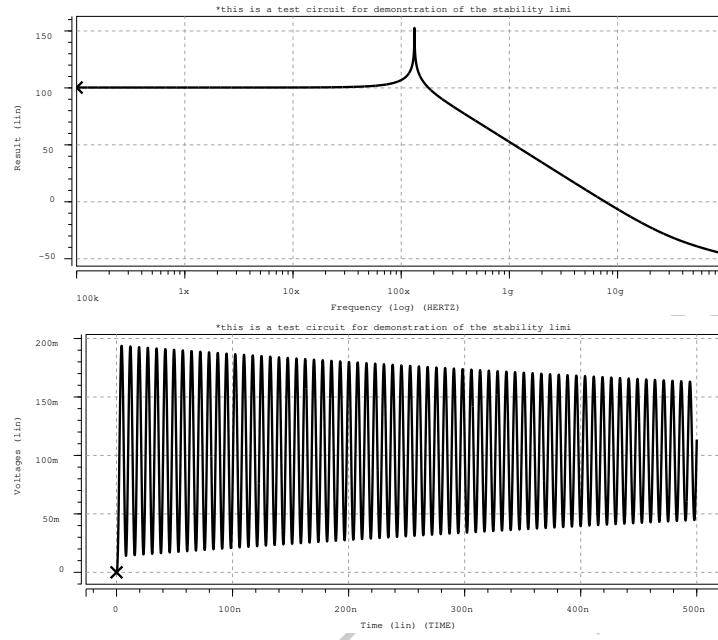


Figure 6.6: The closed-loop transfer function and the step response of the three-stage amplifier of Figure 6.3, with  $R_F = 116\text{k}\Omega$ .

to the gate of  $M_1$ , i.e.,

$$\mathbb{T}(0) = g_{m1}g_{m2}g_{m3}R_1R_2R_3 \cdot \frac{R_S}{R_S + R_F + r_{m4}} \quad (6.3)$$

It is easy to see that there are no zeros in the loop gain since shorting any of the capacitors results in no signal at  $v_y$ . Also it is easy to see that the time constants are uncoupled from each other and thus the loop gain can be stated as:

$$\mathbb{T}(s) = \frac{\mathbb{T}(0)}{(1 + \tau_{in}s)(1 + \tau_1s)(1 + \tau_2s)} \quad (6.4)$$

where

$$\tau_{in} = [R_S \parallel (R_F + r_{m4})]C_{in}$$

$$\tau_1 = R_1C_1$$

$$\tau_2 = R_2C_2$$

which has three real LHP poles. Its Bode plot is shown in FIXXX.

For  $R_F = 112\text{k}\Omega$  we have  $\mathbb{T}(0) = 8.81$ ,  $\tau_{in} = 1.49\text{nS}$ ,  $\tau_1 = 2.5\text{nS}$ ,  $\tau_2 =$

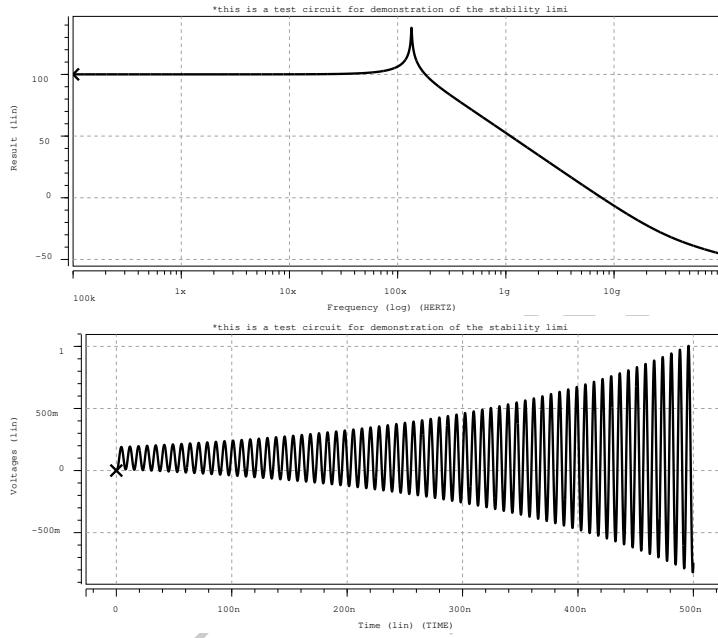


Figure 6.7: The closed-loop transfer function and the step response of the three-stage amplifier of Figure 6.3, with  $R_F = 112\text{k}\Omega$ .

$2.5nS$ ,

$$1 + \mathbb{T}(s) = 0 \Rightarrow \tau_{in} \tau_1 \tau_2 s^3 + (\tau_{in} \tau_1 + \tau_{in} \tau_2 + \tau_1 \tau_2) s^2 + (\tau_{in} + \tau_1 + \tau_2) s + 1 + \mathbb{T}(0) = 0 \quad (6.5)$$

$$9.313(nS)^3 s^3 + 13.7(nS)^2 s^2 + 6.49(nS)s + 9.81 = 0 \quad (6.6)$$

**Example 6.0.2 (Conditional Stability)** Let us go back to the three-stage shunt-shunt trans-impedance amplifier of Figure 6.3, where this time capacitors  $C_1$  and  $C_2$  have series resistances  $R_{z1}$  and  $R_{z2}$ , respectively, as illustrated in Figure 6.10. These resistors could represent the inevitable gate resistances of the MOS capacitors, or explicit resistors in the circuit. In this example, we will investigate how this seemingly small change can result in some interesting behavior in the circuit.

Assuming  $R_{z1} = R_{z2} = 300\Omega$  and under a relatively small feedback resistor (large feedback)  $R_F$  of  $1\text{k}\Omega$ , we can break the loop at the gate of  $M_4$  (still ignoring its capacitors) to determine the loop gain. We can simulate the loop gain under these conditions to obtain the amplitude and phase frequency response of Figure 6.11, shown as Bode plot. We notice that the phase goes below  $180^\circ$  at

#### ♦ Numerical Example ♦

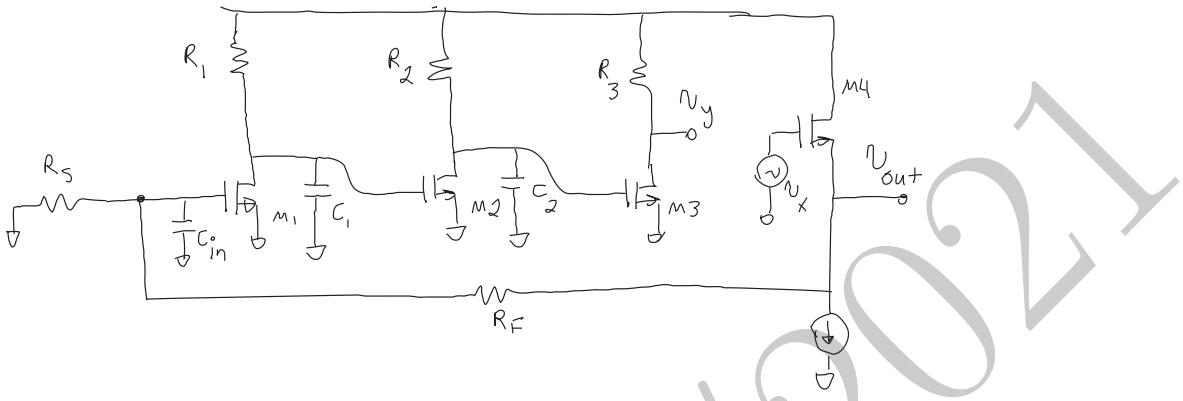


Figure 6.8: Breaking the loop of the shunt-shunt feedback amplifier of Figure 6.3 at the gate of  $M_4$  in the absence of an external input.

292MHz to go below it and then comes back up and crosses  $180^\circ$  one more time at 603MHz. We also notice that at both these frequencies the magnitude of loop gain is greater than unity (positive in dB). This implies that in the broken loop, applying a sinusoidal input,  $v_x$ , at either of these two frequencies will result in a sinusoidal output,  $v_y$ , that will be in-phase (phase difference of an integer multiple of  $360^\circ$  around the loop) and larger in magnitude. We can verify this in simulations by applying these sine waves at either frequencies and observing the output waveforms, as shown in Figures 6.12 and Figure 6.13 for 292MHz and 603MHz, respectively. As we can see in both cases,  $v_y$  is greater than the excitation voltage  $v_x$  and is in phase. Both of these are clear examples of a loop which returns the signal in phase and with an amplitude greater than unity. We saw earlier in Example 6.0.1 that a similar case resulted in a step response that exhibited an exponentially growing sinusoidal at the output of the closed-loop system. Let us evaluate the step response of the system when the loop is closed.

Figure 6.14 shows the step response of the closed loop circuit of Figure 6.10. Somewhat unexpectedly, it shows a decaying sinusoidal step response despite two points with a loop gain which has phase of  $-180^\circ$  and amplitude greater than unity at two points! This is clear example where we have to be careful about the “intuitive” view regarding stability.

The loop gain of the modified circuit of Figure 6.10 can be easily evaluated by noting the discussion in Example 4.4.3 of subsection 4.4.2 in Chapter 4. We noted that when we have a series RC network of  $R_{z1}$  and  $C_{z1}$  in parallel with a load resistor  $R_1$  (Figure 4.46), the transfer function exhibits a pole and a zero given by (4.52). Noting that the time constant for  $C_{in}$  is still uncoupled and unchanged, applying (4.52) to  $C_1$  and  $C_2$ , we can easily determine the loop gain to be:

$$\mathbb{T}(s) = \mathbb{T}(0) \cdot \frac{(1 + \tau_{z1}s)(1 + \tau_{z2}s)}{(1 + \tau_{in}s)(1 + \tau'_1s)(1 + \tau'_2s)} \quad (6.7)$$

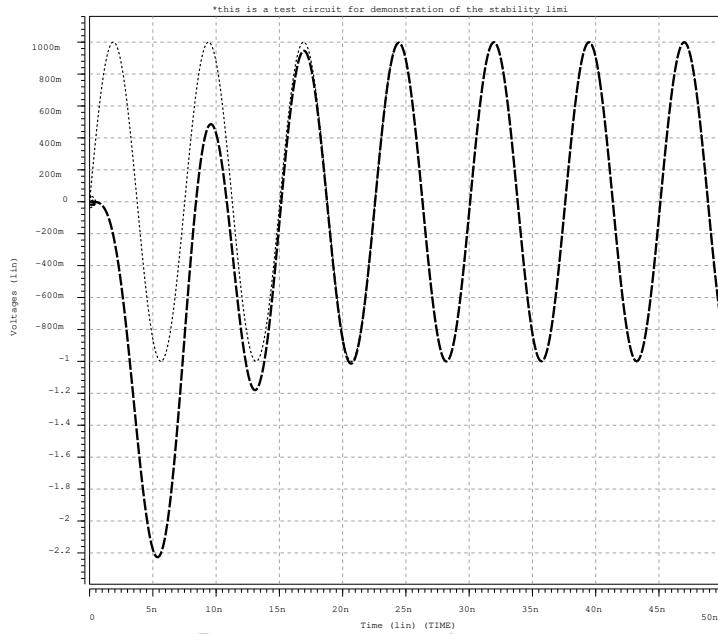


Figure 6.9: The  $v_x$  and  $v_y$  waveform for the broken loop shunt-shunt feedback amplifier of Figure 6.8 where  $v_x$  is sine wave with a frequency of 133MHz.

where

$$\tau_{in} = [R_S \parallel (R_F + r_{m4})]C_{in}$$

$$\tau'_1 = (R_1 + R_{z1})C_1$$

$$\tau_{z1} = \tau'_1 \cdot \frac{\mathbb{T}^{z1}}{\mathbb{T}^0} = R_{z1}C_1$$

$$\tau'_2 = (R_2 + R_{z2})C_2$$

$$\tau_{z2} = \tau'_2 \cdot \frac{\mathbb{T}^{z2}}{\mathbb{T}^0} = R_{z2}C_2$$

where  $\mathbb{T}^0 \equiv \mathbb{T}(0)$ . Alternatively the loop gain can be expressed as:

$$\mathbb{T}(s) = \mathbb{T}(0) \cdot \frac{(1 + R_{z1}C_1s)(1 + R_{z2}C_2s)}{[1 + [R_S \parallel (R_F + r_{m4})]C_{in}s][1 + (R_1 + R_{z1})C_1s][1 + (R_2 + R_{z2})C_2s]} \quad (6.8)$$

with three real LHP poles and two real LHP zeros.

$Z_\infty(s) = -R_F$  and  $Z_0(0) = R_S \cdot \frac{r_{m4}}{R_S + R_F}$  therefore  $Z_0(s) = \frac{Z_0(0)}{1 + \tau'_{in}s}$  with  
 $\tau'_{in} = [R_S \parallel (R_F + r_{m4})]C_{in}$

Calculating the poles of the closed loop system

$$1 + \mathbb{T}(s) = 1 + \mathbb{T}(0) \frac{(1 + \tau_{z1}s)(1 + \tau_{z2}s)}{(1 + \tau_{in}s)(1 + \tau'_1s)(1 + \tau'_2s)} = 0 \quad (6.9)$$

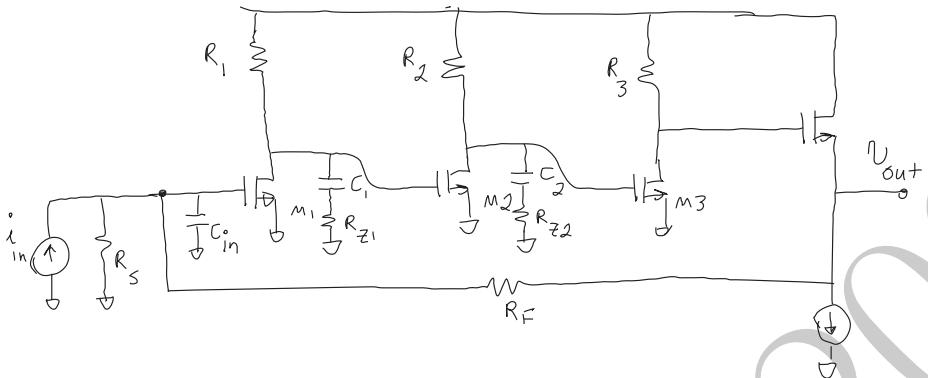


Figure 6.10: The modified three-stage shunt-shunt amplifier of Figure 6.3 where  $C_1$  and  $C_2$  have series resistances  $R_{z1}$  and  $R_{z2}$ , respectively.

thus

$$\tau_{in}\tau'_1\tau'_2s^3 + [\tau_{in}\tau'_1 + \tau'_1\tau'_2 + \tau_{in}\tau'_2 + \tau_{z1}\tau_{z2}\mathbb{T}(0)]s^2 + [\tau_{in} + \tau'_1 + \tau'_2 + (\tau_{z1} + \tau_{z2})\mathbb{T}(0)]s + 1 + \mathbb{T}(0) = 0 \quad (6.10)$$

For  $R_F = 1k\Omega$  we have  $\mathbb{T}(0) = 400$ ,  $\tau_{in} = 0.9nS$ ,  $\tau_1 = 2.65nS$ ,  $\tau_2 = 2.65nS$ ,  $\tau_{z1} = 0.75nS$ ,  $\tau_{z2} = 0.75nS$ ,

$$6.32(nS)^3s^3 + 236.8(nS)^2s^2 + 606.2(nS)s + 401 = 0 \quad (6.11)$$

## 6.1 Nyquist Stability Criterion

Online YouTube lecture:

[Stability criteria: Routh-Hurwitz, Nyquist derivation.](#)

▼ Derivation ▼

Consider that the complex frequency  $s$  traverses a closed contour  $C_1$  in the complex plane, as shown in Figure 6.15a. We can evaluate an arbitrary function of  $s$ , namely  $F(s)$ , and plot its values in the complex plane shown as contour  $C_2$  in Figure 6.15b for the values  $s$  assumes as it traverses the contour  $C_1$ . The resultant contour  $C_2$  will be closed for a closed  $s$  contour  $C_1$  since  $s$  goes back to the same value it started with. For a given complex value (point) of  $s$  on  $C_1$ , the  $F(s)$  will assume a complex value with a magnitude  $|F(s)|$  and a phase angle,  $\angle F(s)$ . It is easy to see that  $\angle F(s)$  is the sum of the angles of the vectors from the zeroes to the current value of  $s$  on the contour  $C_1$  (denoted by  $\theta$ 's)

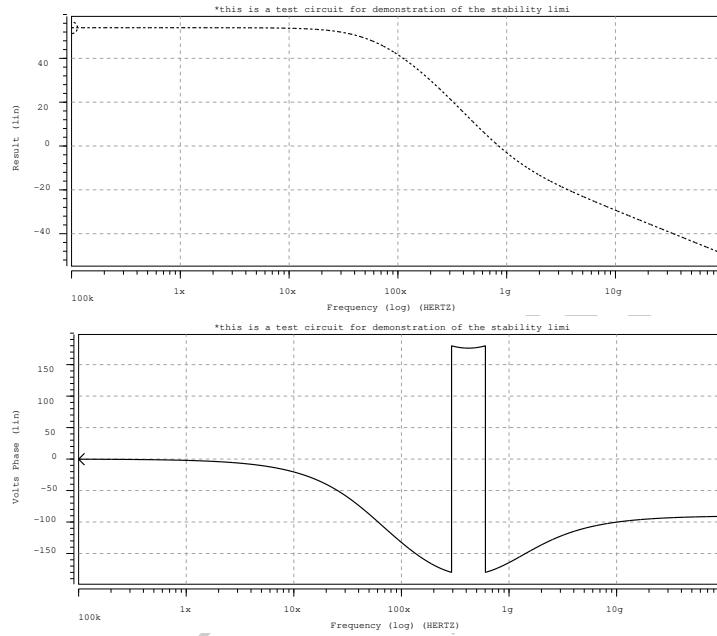


Figure 6.11: The magnitude and phase of the loop gain of the modified three-stage amplifier of Figure 6.10, with  $R_F = 1\text{k}\Omega$  and  $R_{z1} = R_{z2} = 300\Omega$ .

minus the angles of vectors from the poles to the same point (denoted by  $\phi$ 's)<sup>3</sup>.

It is easy to see that if there are no poles or zeros inside the closed contour  $C_1$ , the phase of  $F(s)$  does not experience a  $2\pi$  ( $360^\circ$ ) phase shift as it goes through the contour or equivalently the closed contour  $C_2$  of  $F(s)$  does not encircle the origin in Figure 6.15b. Conversely, it is easy to see that if there are poles and/or zeros inside a clockwise contour  $C_1$ , the number of times the contour  $C_2$  encircles the origin in a clockwise direction is simply equal to the number of zeros minus the number poles inside the contour  $C_1$ , as shown in

<sup>3</sup>To see this note that

$$\begin{aligned} F(s) = |F(s)|e^{j\angle F(s)} &= a \frac{(s - z_1)(s - z_2) \cdots (s - z_m)}{(s - p_1)(s - p_2) \cdots (s - p_m)} \\ &= |a| \frac{|s - z_1||s - z_2| \cdots |s - z_m|}{|s - p_1||s - p_2| \cdots |s - p_m|} \cdot \frac{e^{j\angle(s-z_1)}e^{j\angle(s-z_2)} \cdots e^{j\angle(s-z_m)}}{e^{j\angle(s-p_1)}e^{j\angle(s-p_2)} \cdots e^{j\angle(s-p_m)}} \end{aligned}$$

Therefore

$$\begin{aligned} \angle F(s) &= \angle(s - z_1) + \angle(s - z_2) + \cdots + \angle(s - z_m) - [\angle(s - p_1) + \angle(s - p_2) + \cdots + \angle(s - p_n)] \\ &= \theta_1 + \theta_2 + \cdots + \theta_m - [\phi_1 + \phi_2 + \cdots + \phi_n] \end{aligned}$$

where we note that  $s - z_i$  and  $s - p_i$  are the vectors connecting  $z_i$  and  $p_i$  to point  $s$ , and hence  $\angle(s - z_i)$  and  $\angle(s - p_i)$  are the angles these vectors make with the real axis, respectively.

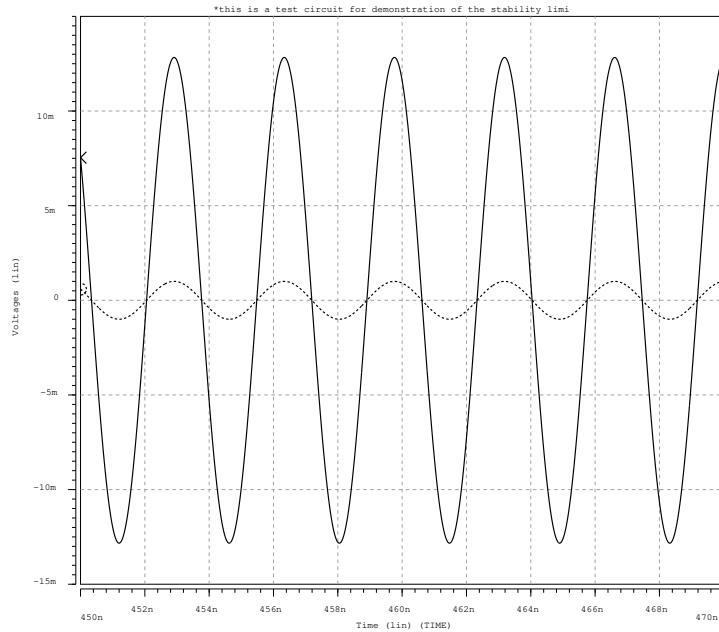


Figure 6.12: The input and the output voltage for the broken loop of the modified three-stage amplifier of Figure 6.10 at 292MHz with  $R_F = 1k\Omega$  and  $R_{z1} = R_{z2} = 300\Omega$ . The loop gain is  $\mathbb{T} \approx 12.8$ .

Figure 6.16. This is known as Cauchy's principle.

According to 5.101 of section 5.4.2,  $1 + \mathbb{T}(s)$  (one plus the loop gain) and  $\Delta$  (the determinant of the  $Y$  matrix) have the same roots. However, from 2.23 of Chapter 2 we know that roots of  $\Delta$  are the poles of the system. So if we can determine whether  $1 + \mathbb{T}(s)$  has RHP roots, we know that we have RHP poles and thus the system will be BIBO unstable.

To determine whether the closed-loop transfer function has RHP poles, we can choose the  $C_1$  contour shown in Figure 6.17a such that it encompasses the entire RHP in the  $s$ -plane. This contour simply corresponds to varying  $s$  from  $-j\infty$  to  $+j\infty$  along the imaginary axis<sup>4</sup>. Then, we evaluate the  $C_2$  contour of the function  $1 + \mathbb{T}(s)$ , where the number of clockwise encirclements of the origin by its trajectory (contour  $C_2$ ) as  $s$  goes from  $-j\infty$  to  $+j\infty$  determines the number of RHP zeros (roots) of  $1 + \mathbb{T}(s)$  minus its RHP poles (e.g., Figure 6.17b). It is easy to see that the poles of  $1 + \mathbb{T}(s)$  are the same as the original

<sup>4</sup>For all practical systems, the transition from  $+j\infty$  back to  $-j\infty$  through infinity will correspond to a single point on the  $C_2$  contour since in all real systems the loop gains diminish to zero at infinite frequency.

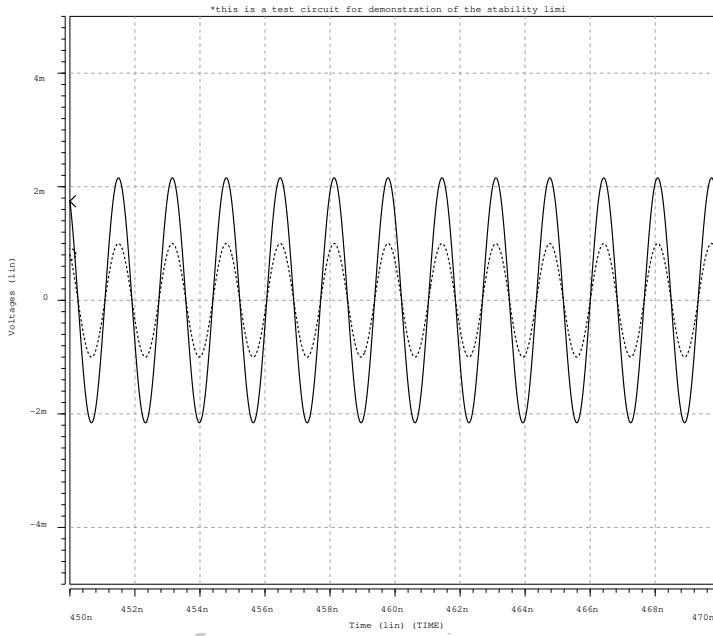


Figure 6.13: The input and the output voltage for the broken loop of the modified three-stage amplifier of Figure 6.10 at 603MHz with  $R_F = 1k\Omega$  and  $R_{z1} = R_{z2} = 300\Omega$ . The loop gain is  $\mathbb{T} \approx 2.2$

poles of the loop gain itself<sup>5</sup>. Therefore, if the original open loop gain  $\mathbb{T}(s)$  is stable and does not have a RHP poles, neither will  $1 + \mathbb{T}(s)$ . So for an open-loop stable system, the number times of  $1 + \mathbb{T}(s)$  encircles the origin simply determines the number of zeros of the  $1 + \mathbb{T}(s)$  which is equal to the number of RHP poles the closed-loop system has.

We note that encirclement of the origin by one plus the loop gain,  $1 + \mathbb{T}(s)$ , is simply equivalent to encirclement of point  $-1$  by the loop gain,  $\mathbb{T}(s)$ . It is easier to plot  $\mathbb{T}(s)$  once we determine the loop gain. Also since for a real system  $\mathbb{T}(s)$  is a transfer function with real coefficients, the contour  $C_2$  evaluated from  $s \rightarrow -j\infty$  to  $s = 0$  is simply the mirror image of the one evaluated from  $s = 0$  to  $s \rightarrow +j\infty$ . Thus, it is sufficient to look at the contour of  $\mathbb{T}(s)$  from  $s = 0$  to  $s \rightarrow +j\infty$ , as shown in Figure 6.17c, which is simply the range of input

<sup>5</sup>To see this we can write  $\mathbb{T}(s) = N(s)/D(s)$ , where  $N(s)$  and  $D(s)$  are the numerator and the denominator of the loop gain. Then it is easy to see that

$$1 + \mathbb{T}(s) = 1 + \frac{N(s)}{D(s)} = \frac{D(s) + N(s)}{D(s)}$$

where the poles of  $\mathbb{T}(s)$  and  $1 + \mathbb{T}(s)$  are the same.

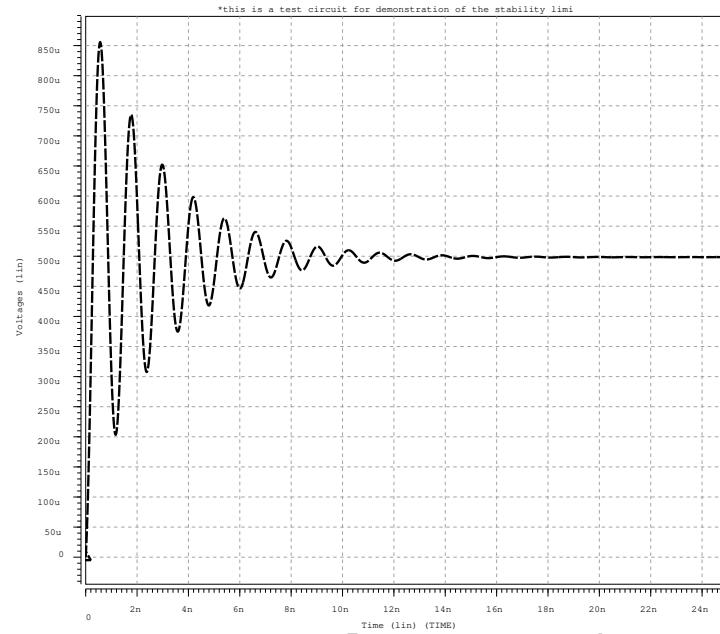


Figure 6.14: The step response of the closed-loop modified three-stage amplifier of Figure 6.10 with  $R_F = 1k\Omega$  and  $R_{z1} = R_{z2} = 300\Omega$ .

frequencies used in a Bode plot.

### ▼ Result ▼

Nyquist stability criterion states that for an LTI system with a stable open loop gain of  $T(s)$ , the number of time the contour of  $T$  in the complex plane<sup>6</sup> encircles the point  $-1$  is equal to the number of RHP poles of the closed-loop system. In other words, starting with an open-loop stable LTI system, the closed-loop system will be stable if the contour of  $T(s)$  does not encircle the point  $-1$  on the real axis.

## 6.2 Simplified Intuitive View

Consider the idealized feedback amplifier of Figures 5.36a, where the loop could be broken at different places. In a real circuit one has to be careful about the source and load impedances, as discussed in great details in Section 5.4, thus this breaking of the loop looks more like Figure 5.37. For our discussion here we will use the simplified picture and will not show these source and load impedances explicitly for the time being.

Let us say in the broken feedback loop of FIXXX, we introduce an periodic

<sup>6</sup>i.e.,  $Im[T(s)]$  vs.  $Re[T(s)]$  as  $s$  varies from 0 to  $+\infty$ .

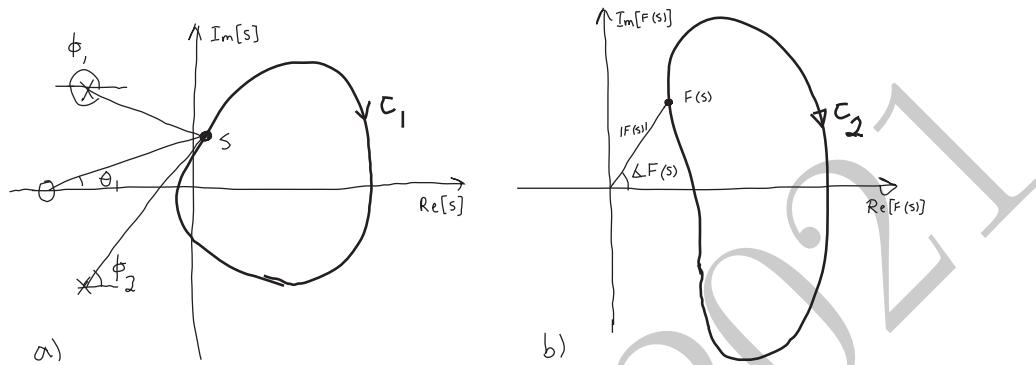


Figure 6.15: a) A closed contour of  $s$  in the complex plane, b) the corresponding trajectory of the transfer function  $F(s)$  for the range of values  $s$  assumes in contour  $C_1$ .

excitation waveform into the system and observe the returned signal to determine the loop gain (under appropriate loading conditions). Generally speaking the returned signal will have a different shape and size.

Now consider the special case of a periodic waveform that goes through the system is returned *exactly* with the same shape and amplitude with exactly enough *delay*<sup>7</sup> that it coincides with one of subsequent cycles of the periodic input waveform<sup>8</sup>. In that case, if the loop is instantaneously closed, it can maintain this waveform without external intervention. In other word, such a periodic waveform can be a steady-state solution of the system.

Now imagine the scenario where the returned signal has exactly the same amplitude as the the excitation. In this case, the system *may* maintain this waveform indefinitely and oscillate. There are other conditions that are necessary for the system to actually have stable oscillations that will be discussed in XXX.

For the special case when the system of FIXXX is LTI, the loop gain as a function of frequency can be measured by using single-frequency (i.e., lasting a long time) sinusoidal inputs in the setup and monitoring the phase and the amplitude of the returned signal<sup>9</sup>. This result can be plotted in different ways such as the Bode plot, as shown in FIXXX.

For a discussion of non-linear stability criteria, watch online YouTube lecture:  
[\*\*Nonlinear stability criteria: Circle criterion, off-axis circle criterion\*\*](#)

A brief overview of phase and gain margin can be found in the online YouTube lecture:

<sup>7</sup>As it will become evident later in XXX, the distinction between phase and delay plays an important role in stability analysis of the system.

<sup>8</sup>i.e., the delay is an integer multiple of the fundamental frequency of the periodic signal.

<sup>9</sup>Again, we are assuming that we have properly taken care of the loading effects.

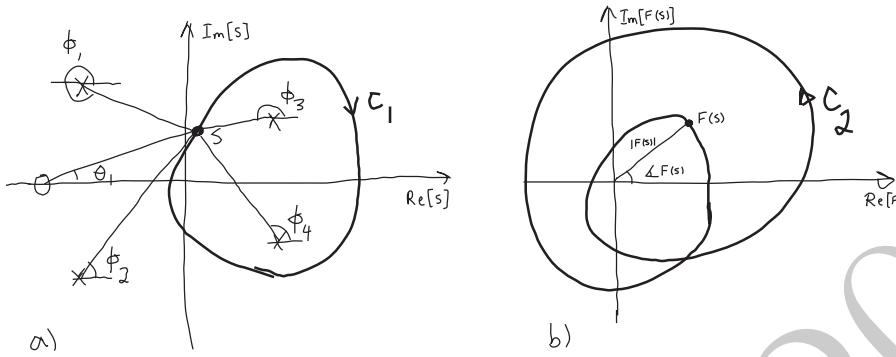


Figure 6.16: a) A closed contour of  $s$  in the complex plane that encompasses two poles, b) the corresponding trajectory of the transfer function  $F(s)$  for the range of values  $s$  assumes in contour  $C_1$  of part a) that encircles the origin twice.

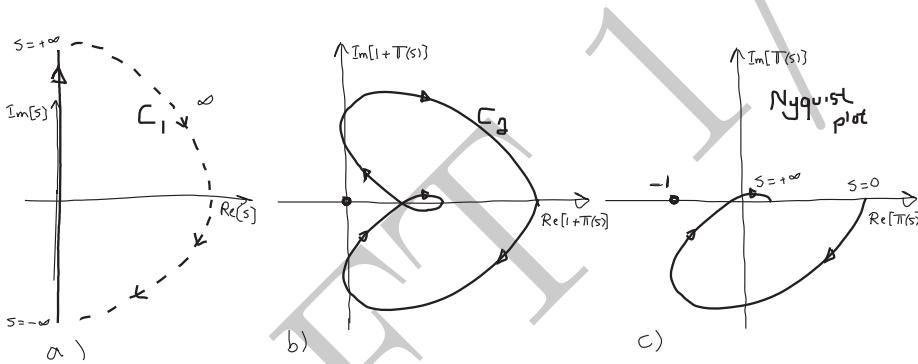


Figure 6.17: a) The contour along the  $j\omega$  axis used for Nyquist stability plot, b) the trajectory of  $1 + T$ , c) half-plane ( $s$  from 0 to  $+j\infty$ ) trajectory of  $T$ .

### Phase margin and gain margin

A review of compensation principle and its application is presented in the following online YouTube lectures:

[Feedback amplifier compensation, general view for 1st, 2nd, and 3rd order system](#)  
[Circuit compensation techniques, one- and two-stage op-amp, Miller compensation](#)

DRAFT 1/2021