Ecole Royale Militaire

Bruxelles — Belgique

ES222
Electronique

Koen Boeckx

April 5, 2023

Contents

Chapter 1

Introduction

Ce cours est la suite du cours d'électronique donné par le Prof. dr. ir. Patrick MERKEN, qui a créé le contenu du cours et conçu la plupart des figures utilisées dans le travail. Mes remerciements vont à lui.

1.1 Aperçu du cours

Au début du cours, nous couvrirons les fondamentaux de la théorie des circuits, qui incluent les composants passifs tels que les résistances, les inductances et les capacités, ainsi que les lois de Kirchhoff, le comportement en fréquence des systèmes linéaires et la transformation de Thévenin et Norton. Nous supposons que les étudiants ont une compréhension de base de ces concepts.

La première grande section du cours, intitulée *Composants*, introduira les étudiants aux bases de la physique des semi-conducteurs. Cela nous permettra de développer une compréhension des diagrammes de bande et des semi-conducteurs. Nous explorerons ensuite deux dispositifs à semi-conducteurs courants, à savoir les diodes et les transistors. Les deux types de transistors, à jonction bipolaire (BJTs) et à effet de champ à grille isolée en oxyde métallique (MOSFETs), seront couverts en détail.

Ensuite, nous passerons à la deuxième section du cours, intitulée *Électronique analogique*. Ici, nous approfondirons le comportement des composants électroniques dans les circuits. Nous étudierons en détail les amplificateurs et les amplificateurs opérationnels (op-amps), ainsi que les oscillateurs, les références de tension et la théorie de la rétroaction. Un concept important que nous introduirons dans cette section est le modèle de petit signal, qui est utilisé pour étudier le comportement d'un circuit autour d'un point de fonctionnement.

Enfin, nous couvrirons la troisième section du cours, intitulée *Électronique numérique*. Ici, nous explorerons comment les transistors peuvent être utilisés comme des interrupteurs pour construire des portes logiques pour effectuer des calculs. Nous discuterons également des circuits séquentiels qui contiennent des éléments de mémoire, ainsi que de la transition entre les domaines analogique et numérique. Cela comprendra une discussion sur les convertisseurs analogique-numérique et numérique-analogique.

1.2 Une brève histoire de l'électronique

L'électronique est un domaine qui a connu une évolution spectaculaire au cours du siècle dernier, avec ses origines ancrées dans les découvertes liées à l'électricité et au comportement des charges électriques. L'histoire de l'électronique est un voyage fascinant à travers les découvertes scientifiques, les innovations technologiques et l'évolution des dispositifs électroniques qui ont transformé notre façon de vivre et de travailler.

1.2.1 Les premières découvertes et innovations

Les racines de l'électronique remontent à la découverte de l'électricité par Benjamin Franklin au XVIIIe siècle. Cependant, ce n'est qu'au XIXe siècle que plusieurs découvertes et innovations clés ont préparé le terrain pour le développement de l'électronique moderne.
L'une des découvertes les plus importantes a été faite par Michael Faraday, qui a démontré au début des années 1800 qu'un champ magnétique changeant pouvait induire un courant électrique dans un fil à proximité. Ce phénomène, connu sous le nom d'induction électromagnétique, a ouvert la voie au développement de générateurs et de moteurs, qui sont devenus des composants essentiels de nombreux dispositifs électroniques.
En 1873, James Clerk Maxwell a publié un ensemble d'équations décrivant le comportement des champs électriques et magnétiques :

$$\nabla \cdot \vec{E} = \frac{\rho}{\epsilon} \quad \nabla \cdot \vec{B} = 0 \quad \nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \quad \nabla \times \vec{B} = \mu \vec{J} + \epsilon\mu \frac{\partial \vec{E}}{\partial t}$$

Ces équations ont unifié les théories de l'électricité et du magnétisme et ont prédit l'existence d'ondes électromagnétiques, qui se déplacent à la vitesse de la lumière. Cette découverte a été déterminante dans le développement de la communication radio, qui allait plus tard révolutionner notre manière de communiquer.

1.2.2 Les premiers dispositifs électroniques

Les premiers dispositifs électroniques ont émergé à la fin du XIXe siècle. En 1904, John Ambrose Fleming a inventé le tube à vide, qui était un type de tube électronique pouvant être utilisé comme amplificateur ou commutateur. Le tube à vide était un composant critique dans les premiers radios et télévisions et était la technologie principale utilisée pour l'amplification électronique jusqu'au développement du transistor au milieu du XXe siècle.
En 1907, Lee De Forest a inventé le triode, un type de tube à vide capable d'amplifier des signaux électriques en contrôlant le flux d'électrons à travers le vide. Le triode a rendu possible l'amplification de signaux avec une grande précision, en faisant un composant essentiel des premiers radios et autres dispositifs électroniques.

1.2.3 L'avènement de l'électronique moderne

Le développement du transistor en 1947 a marqué un tournant significatif dans l'histoire de l'électronique. Le transistor, inventé par John Bardeen, Walter Brattain et William Shockley chez Bell Labs, était un dispositif à semi-conducteurs qui pouvait être utilisé comme commutateur ou amplificateur. Les transistors étaient plus petits, plus rapides et plus fiables que les tubes électroniques et ont permis la miniaturisation des dispositifs électroniques.

Le développement du circuit intégré dans les années 1950 et 1960 a encore révolutionné l'électronique en permettant à plusieurs transistors et autres composants d'être fabriqués sur une seule pièce de silicium. Cela a permis la création de systèmes électroniques complexes qui pouvaient être beaucoup plus petits, plus légers et plus abordables que jamais auparavant.
Au cours des dernières décennies, l'électronique a continué à évoluer avec le développement de nouvelles technologies telles que les microprocesseurs, le traitement numérique du signal et la communication sans fil. Aujourd'hui, l'électronique joue un rôle essentiel dans presque tous les aspects de la vie moderne, des smartphones et des ordinateurs aux dispositifs médicaux et aux technologies d'énergie renouvelable.

Chapter 2

Théorie des Circuits

Dans ce chapitre, nous réitérons plusieurs éléments de la théorie des circuits élémentaires qui sont nécessaires pour comprendre les circuits électroniques de base. Ces éléments comprennent des composants passifs tels que les résistances, les condensateurs et les bobines, les lois de Kirchhoff, la représentation de phasor et l'impédance complexe, ainsi que plusieurs transformations de circuit pour faciliter l'analyse.

2.1 Composants Passifs

Dans les circuits électroniques que nous étudierons, nous rencontrons quatre types d'éléments :

- Sources : ce sont des systèmes relativement complexes avec deux bornes qui fournissent soit une tension, soit un courant. Les sources peuvent être constantes, c'est-à-dire que le courant ou la tension générée ne varie pas avec le temps, ou elles peuvent générer un courant ou une tension qui varie dans le temps. Dans le premier cas, nous les appelons sources de courant continu (DC), dans le second cas, nous parlons de sources de courant alternatif (AC), généralement avec une valeur moyenne de zéro.
- Éléments linéaires : ceux-ci peuvent être passifs, tels que les résistances, les condensateurs ou les bobines, ou actifs, tels que les sources de courant ou de tension dépendantes. Ces dernières sont des sources qui dépendent linéairement d'autres courants ou tensions dans le circuit.
- Éléments non linéaires, tels que les diodes et les transistors. L'étude de l'électronique concerne ces éléments et la façon dont ils sont utilisés dans les circuits.
- Conducteurs, qui connectent les différents éléments discrets. Nous supposons qu'ils sont idéaux : ils n'ont pas de résistance, d'inductance ou de capacité. Si l'un de ces défauts est présent dans des conducteurs réels, ils seront modélisés comme des éléments discrets séparés.

Nous utilisons toujours l'approximation quasi-statique, où la valeur du courant dans une branche est la même partout dans cette branche. Ceci est valable lorsque les dimensions du circuit sont beaucoup plus petites que la longueur d'onde du signal.
The most important linear passive components are:

- A *resistor* is an element that resists the flow of current due to an applied voltage. The current-voltage relation is:

$$v_R = R\, i_R$$

The resistance of a resistor is measured in ohms (Ω).

- A *capacitor* is a passive electronic component that stores electrical energy in an electric field. It consists of two conductive plates separated by an insulating material, called the dielectric.
When a voltage is applied across the plates of a capacitor, electrical charge Q accumulates on the plates, creating an electric field between them. The current-voltage relation is

$$v_C = \frac{Q}{C} = \frac{1}{C}\int i_C\, dt$$

or

$$i_C = C\,\frac{dv_C}{dt}$$

A capacitor resists a sudden change in voltage across its terminals. The capacitance of a capacitor is measured in farad (F).

- An *inductor* is a passive electronic component that stores electrical energy in a magnetic field. It consists of a coil of wire, often wrapped around a core made of a magnetic material, such as iron or ferrite.
When an electric current flows through an inductor, a magnetic field is created around the coil. The strength of the magnetic field is proportional to the amount of current flowing through the coil. When the current changes, the magnetic field changes, inducing a voltage across the coil that opposes the change in current. Thus, an inductor resists a sudden change in current. The current-voltage relation is

$$v_L = L\,\frac{di_L}{dt}$$

or also

$$i_L = \frac{1}{L}\int v_L\, dt$$

The inductance of an inductor is measured in henry (H).

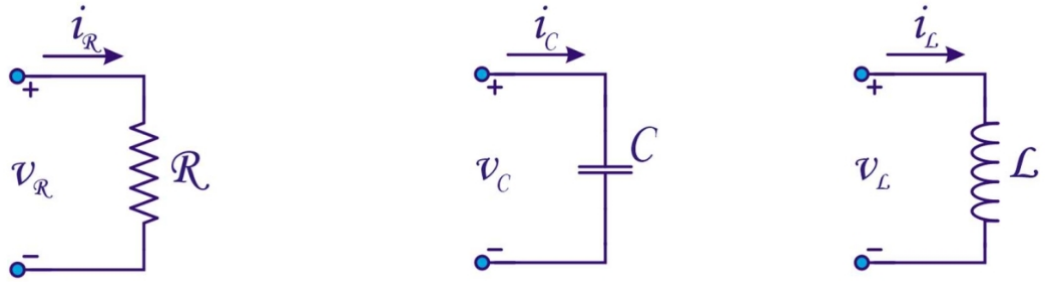


Figure 2.1: Resistor (left), capacitor (middle) and inductor (right)

Les composants passifs linéaires les plus importants sont :

- Une *résistance* est un élément qui résiste à l'écoulement du courant en raison d'une tension appliquée. La relation courant-tension est :

$$v_R = R\, i_R$$

La résistance d'une résistance est mesurée en ohms (Ω).

- Un *condensateur* est un composant électronique passif qui stocke de l'énergie électrique dans un champ électrique. Il se compose de deux plaques conductrices séparées par un matériau isolant appelé diélectrique.
Lorsqu'une tension est appliquée aux plaques d'un condensateur, une charge électrique Q s'accumule sur les plaques, créant un champ électrique entre elles.

La relation courant-tension est

$$v_C = \frac{Q}{C} = \frac{1}{C}\int i_C\, dt$$

ou encore

$$i_C = C\,\frac{dv_C}{dt}$$

Un condensateur résiste à un changement soudain de tension à ses bornes. La capacité d'un condensateur est mesurée en farads (F).

- Une *inductance* est un composant électronique passif qui stocke de l'énergie électrique dans un champ magnétique. Elle se compose d'une bobine de fil, souvent enroulée autour d'un noyau en un matériau magnétique tel que le fer ou la ferrite. Lorsqu'un courant électrique circule dans une inductance, un champ magnétique est créé autour de la bobine. La force du champ magnétique est proportionnelle à la quantité de courant circulant dans la bobine. Lorsque le courant change, le champ magnétique change, ce qui induit une tension à travers la bobine qui s'oppose au changement de courant. Ainsi, une inductance résiste à un changement soudain de courant. La relation courant-tension est

$$v_L = L\,\frac{di_L}{dt}$$

ou encore

$$i_L = \frac{1}{L}\int v_L\, dt$$

L'inductance d'une inductance est mesurée en henry (H).

2.2 Lois de Kirchhoff

2.2.1 Loi des tensions de Kirchhoff

La loi des tensions de Kirchhoff (LTV) stipule que la somme des différences de potentiel (tensions) autour d'une boucle fermée est nulle :

$$\sum_{k=1}^{n} v_k = 0 \tag{2.1}$$

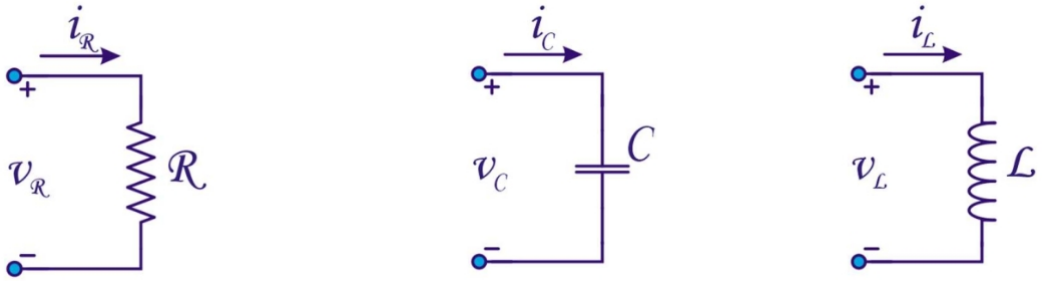


Figure 2.2: Résistance (à gauche), condensateur (au milieu) et inductance (à droite)

La LTV est en réalité une reformulation de la loi de Faraday $\nabla \times \vec{E} = -\frac{\partial B}{\partial t} = 0$, où l'on suppose que les champs magnétiques (variables dans le temps) sont confinés à chaque composant et que le champ dans la région extérieure au circuit est négligeable.

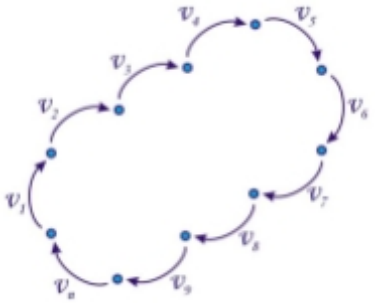


Figure 2.3: Loi des tensions de Kirchhoff

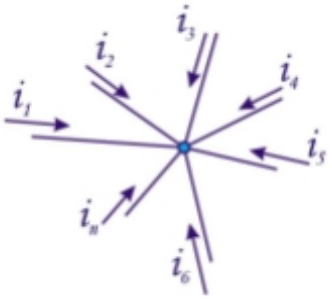


Figure 2.4: Loi des courants de Kirchhoff

2.2.2 Loi des courants de Kirchhoff

La loi des courants de Kirchhoff (LCK) affirme que dans un nœud d'un circuit électrique, la somme algébrique des courants est nulle :

$$\sum_{k=1}^{n} i_k = 0 \tag{2.2}$$

ou de manière équivalente, que la somme des courants entrant dans ce nœud est égale à la somme des courants en sortant. La loi repose sur le fait qu'il n'y a pas d'accumulation de charges dans aucun nœud du réseau.

2.2.3 Exemple

Considérons le circuit de la figure 2.5. Avec KVL, on peut écrire :

$$v_{in} = v_R + v_C = R\, i + v_C$$

avec $i = C; \frac{dv_C}{dt}$. L'équation pour trouver v_C devient :

$$v_{in} = RC\, \frac{dv_C}{dt} + v_C$$

Si v_{in} est soudainement allumé (par exemple, en fermant un interrupteur à $t = 0$), alors $v_{in} = V_0; u(t)$ avec $u(t)$ la fonction échelon. On peut alors résoudre pour $v_C(t)$:

$$\begin{aligned} RC; \frac{dv_C}{dt} + v_C &= V_0 \\ \frac{dv_C}{dt} &= \frac{1}{RC}(V_0 - v_C) \\ \frac{dv_C}{V_0 - v_C} &= \frac{dt}{RC} \\ \int_0^{v_C} \frac{dv_C}{V_0 - v_C} &= \int_0^t \frac{dt}{RC} \\ -\ln(V_0 - v_C) &= \frac{t}{RC} + K' \\ v_C(t) &= V_0 - Ke^{\frac{-t}{RC}} \end{aligned}$$

avec $K = V_0$ de telle sorte que $v_C(t = 0) = 0$:

$$v_C(t) = V_0(1 - e^{\frac{-t}{RC}}) = V_0(1 - e^{-\frac{t}{T}})$$

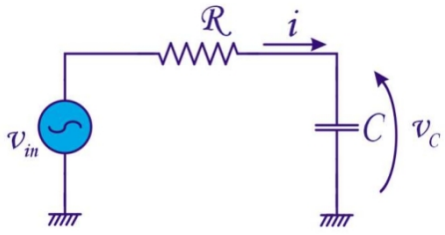


Figure 2.5

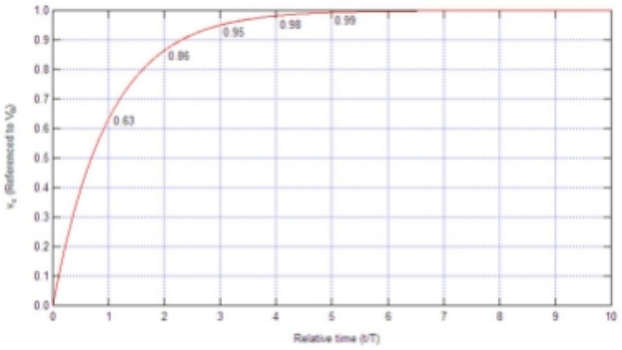


Figure 2.6

Cette fonction est représentée dans la figure 2.6. La tension v_C augmente comme la réponse à un échelon d'un système du premier ordre, et atteint 63% de la valeur finale après un temps caractéristique $T = RC$. Il faut $5T$ pour atteindre 99% de la valeur finale.

2.3 Représentation en fréquence

2.3.1 La transformation de Steinmetz

Supposons qu'un signal sinusoïdal avec une fréquence $f = \omega/2\pi$ soit appliqué à un circuit ne contenant que des éléments linéaires. La théorie des systèmes linéaires nous enseigne alors que chaque courant et tension sera également une sinusoïde avec la même fréquence. Supposons que le courant dans un élément peut être écrit comme $I = I_0 \cos(\omega; t) = ReI_0 e^{j\omega t}$. Dans ce cas, nous pouvons écrire :

- Pour une résistance : $V_R = R; I_R$
- Pour un condensateur : $V_C = \frac{1}{C} \int I_L \, dt = \frac{1}{C} \int I_0 e^{j\omega t} \, dt = \frac{1}{j\omega C} I_0 e^{j\omega t} = \frac{1}{j\omega C} \, I_C$

- Pour une inductance : $V_L = L\frac{dI_L}{dt} = L\frac{d(I_0 e^{j\omega t})}{dt} = j\omega L; I_0 e^{j\omega t} = j\omega L; I_L$

Ces relations sont la représentation en phaseur d'un signal sinusoïdal, ou les transformations de Steinmetz. Elles sont résumées dans la figure 2.7.

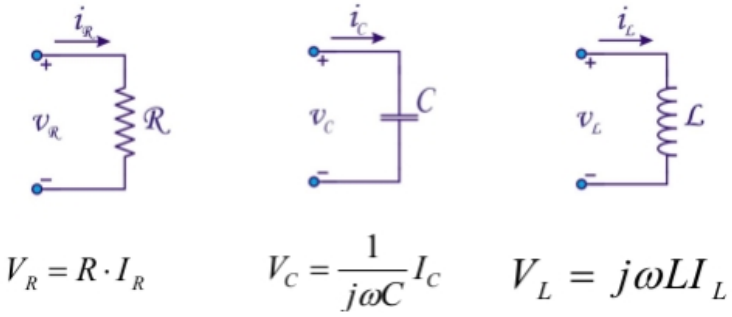


Figure 2.7

De cela, nous concluons que pour chaque élément, nous pouvons écrire:

$$V = Z \cdot I$$

où Z est l'*impédance complexe* de l'élément : R pour une résistance, $\frac{1}{j\omega C}$ pour un condensateur et $j\omega L$ pour une bobine. Nous définissons également d'autres paramètres de circuit : $Y = \frac{1}{Z}$ comme l'*admittance* telle que $I = Y \cdot V$, et $G = \frac{1}{R}$ comme la *conductance* (mesurée en siemens ou A/V).

2.3.2 La transformation de Laplace

La représentation phasorique ou la transformée de Steinmetz n'est possible que pour des signaux sinusoïdaux. Cette condition est parfois trop restrictive et nous avons besoin d'autres méthodes d'analyse. L'une de ces méthodes, qui est valable pour des signaux arbitraires $x(t)$, est la transformée de Laplace (unilatérale) $X(s)$, définie comme :

$$\mathcal{L}[x(t)] = X(s) = \int_{t=0}^{+\infty} e^{-st} x(t) dt$$

avec s la variable de Laplace complexe : $s = \sigma + j\omega$. Comme exemple, calculons $\mathcal{L}[\frac{dx(t)}{dt}]$:

$$\mathcal{L}[\frac{dx(t)}{dt}] = \int_{t=0}^{+\infty} e^{-st}\frac{dx}{dt}dt \quad = e^{-st}x(t)|t = 0^{+\infty} - \int t = -\infty^{+\infty} x(t)\frac{d(e^{-st})}{dt}dt \; = \int_{t=0}^{+\infty} s; x(t)e^{-st}dt - x(0^+$$

Lorsque nous supposons que $x(0^+) = 0$, nous trouvons que $\mathcal{L}[\frac{dx(t)}{dt}] = sX(s)$. Un système linéaire avec une entrée $x(t)$ et une sortie $y(t)$ peut être décrit par une équation différentielle linéaire d'ordre supérieur :

$$y(t) = a_0 x(t) + a_1\frac{dx}{dt} + a_2\frac{d^2x}{dt^2} + \ldots + b_1\frac{dy}{dt} + b_2\frac{d^2y}{dt^2} + \ldots \; \Rightarrow \frac{Y(s)}{X(s)} \quad = \frac{a_0 + a_1 s + a_2 s^2 + \ldots}{1 - b_1 s - b_2 s^2 - \ldots}$$

Ainsi, la transformée de Laplace de tout système linéaire peut être écrite comme le rapport de deux polynômes en s. Supposons que nous avons un système du premier ordre :

$$\frac{dy}{dt} = \alpha; y(t) + \beta; x(t) \tag{2.3}$$

Si il n'y avait pas d'entrée $x(t)$, la solution serait :

$$y(t) = Ce^{\alpha t}$$

Cette sortie reste finie pour $t \to \infty$ uniquement lorsque $\alpha \leq 0$. C'est le critère de stabilité pour un système du premier ordre. Si nous prenons la transformée de Laplace de l'équation 2.3, nous trouvons :

$$sY(s) = \alpha; Y(s) + \beta; X(s) \;\Rightarrow \frac{Y(s)}{X(s)} = \frac{\beta}{s - \alpha}$$

Les racines du dénominateur sont les *pôles* de la transformée de Laplace. Dans ce cas, nous avons un seul pôle en $s = \alpha$. Nous avons déjà déterminé que le système est stable lorsque $\alpha \leq 0$. C'est une règle générale : **un système $S(s)$ est stable si tous ses pôles se trouvent dans le demi-plan gauche du plan complexe (LHP)**, c'est-à-dire si leur partie réelle est négative. Nous utilisons principalement la représentation de Steinmetz car les signaux que nous considérons sont principalement sinusoïdaux. Cependant, dans certains cas, nous aurons besoin de la transformée de Laplace.

2.3.3 Combinaisons en série et en parallèle

Deux impédances sont en série lorsque le même courant I les traverse, comme illustré sur la figure 2.8. La chute de tension sur les deux est donc $Z_1 \cdot I + Z_2 \cdot I = (Z_1 + Z_2) \cdot I = Z \cdot I$ avec $Z = Z_1 + Z_2$ la combinaison en série de Z_1 et Z_2.

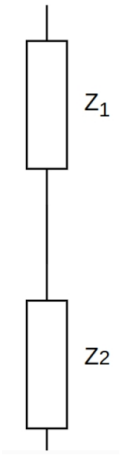


Figure 2.8

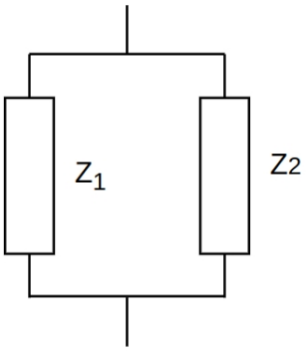


Figure 2.9

Deux impédances sont en parallèle lorsqu'une même tension V est appliquée à leurs bornes, comme illustré sur la figure 2.9. Le courant à travers Z_1 est V/Z_1 et le courant à travers Z_2 est V/Z_2. Le courant total I à travers les deux est donc $V/Z_1 + V/Z_2 = V\left(\frac{1}{Z_1} + \frac{1}{Z_2}\right) = \frac{V}{Z}$ avec $Z = \left(\frac{1}{Z_1} + \frac{1}{Z_2}\right)^{-1}$ la combinaison en parallèle de Z_1 et Z_2.

2.3.4 Théorème de Millman

Le théorème de Millman est dérivé de la loi de Kirchhoff sur les courants et permet de calculer la tension dans un nœud en fonction des tensions dans les nœuds voisins (voir figure 2.10):

$$v_0 = \frac{\sum_{i=1}^{n} Y_i v_i + I_{eq}}{\sum_{i=1}^{n} Y_i} \tag{2.4}$$

avec Y_i la conductance de chaque élément.

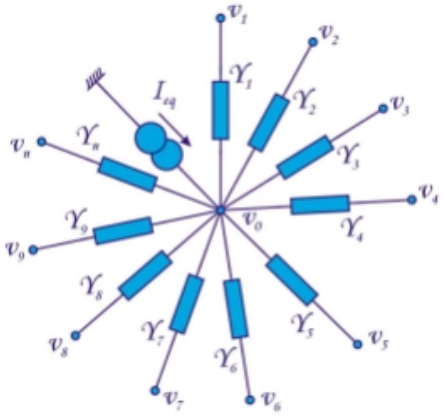


Figure 2.10

On peut déduire ce théorème en appliquant la loi de KCL au noeud 0:

$$\sum_{i=1}^{n} I_n = \sum_{i=1}^{n} Y_i(v_i - v_0) + I_{eq} = 0$$

$$\Rightarrow v_0 \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} Y_i \, v_i + I_{eq}$$

$$\Rightarrow v_0 = \frac{\sum_{i=1}^{n} Y_i v_i + I_{eq}}{\sum_{i=1}^{n} Y_i}$$

La tension dans le noeud 0 est donc la moyenne pondérée des tensions environnantes, avec les conductances comme poids.

2.3.5 Diagramme de Bode

Appliquons les concepts de cette section au problème de la figure 2.5. Comme nous avons supposé que tous les signaux sont sinusoïdaux, nous posons que $v_{in} = V_0 e^{j\omega t}$. Par conséquent:

$$V_{in} = I \cdot R + \frac{1}{j\omega C} I$$

$$\Rightarrow I = \frac{j\omega C}{1 + j\omega RC}; V_{in}$$

$$\text{et} \quad V_C = \frac{1}{1 + j\omega RC}; V_{in}$$

Nous pouvons donc conclure que:

$$\frac{V_C}{V_{in}} = \frac{1}{1 + j\omega RC}$$

$$= \frac{1}{\sqrt{1 + \omega^2 T^2}} e^{-j\phi}$$

avec $\phi = \text{atan}(\omega; T)$. Notez que nous ne trouvons aucune information sur la réponse transitoire comme précédemment, mais nous trouvons comment le circuit se comporte en *régime harmonique permanent* (RHP). Le rapport $\frac{V_C}{V_{in}}$ est appelé *transmittance* $H(\omega)$ et il a une amplitude $A = |T(\omega)|$ et une phase ϕ. Lorsque nous traçons $20\log(A)$ [dB] en fonction de $\log(\omega)$, nous trouvons la courbe de Bode:

$$20\log(A) = -20\,\frac{1}{2}\,\log_{10}(1+\omega^2T^2)$$

À partir de cette expression, nous déduisons que:

- Lorsque $\omega T \ll 1$, $20\log_{10}(A) \approx 20\log_{10}(1) = 0$.
- D'autre part, lorsque $\omega T \gg 1$, $20\log_{10}(A) \approx -20\log_{10}(\omega T)$ et $|H(\omega)|$ diminue de 20 dB pour chaque augmentation de ω d'un facteur de 10 (-20 dB par décade).

En règle générale, toute transmittance $T(\omega)$ peut être écrite sous la forme :

$$T(\omega) = A_0\frac{(1+j\omega T_{n+1})\ldots(1+j\omega T_p)}{(1+j\omega T_1)\ldots(1+j\omega T_n)}$$

Chaque terme individuel est égal à 1 si $\omega \ll 1/T_i$ ou égal à $j\omega T_i$ si $\omega \gg 1/T_i$. Par conséquent, une courbe de Bode peut être construite approximativement en suivant quelques règles simples : lorsque l'on rencontre une pulsation critique $\omega = 1/T$ dans le numérateur (un *zéro*) ou dans le dénominateur (un *pôle*), la courbe va :

- décroître de 20 dB par décade pour un pôle,
- augmenter de 20 dB par décade pour un zéro,

Des règles similaires existent pour la phase ϕ du signal : chaque pôle introduit un décalage de $-\frac{\pi}{2}$, chaque zéro crée un décalage de $+\frac{\pi}{2}$. Notez que ces règles ne produisent qu'une estimation des courbes de Bode ; le tracé réel présentera un comportement plus compliqué, surtout lorsque les pôles sont complexes. Un exemple basé sur la figure 2.5 est donné dans la figure 2.11. Pour un système du premier ordre comme celui-ci, la coupure à -3 dB est atteinte à $\omega = \omega_0$, la pulsation critique.

2.4 Transformations de circuits

2.4.1 Théorème de Thévenin

Le théorème de Thévenin stipule qu'un réseau électrique linéaire à deux bornes peut être remplacé aux bornes A-B par une combinaison équivalente d'une source de tension E_{Th} en série avec une impédance Z_{Th}. Deux circuits sont équivalents s'ils ont la même relation tension-courant à leurs bornes. Cette idée est illustrée dans la figure 2.12, où les éléments du circuit dans le nuage bleu sont remplacés par la source et l'impédance à droite.

- La tension équivalente E_{Th} est la tension obtenue aux bornes A-B du réseau avec les bornes A-B en circuit ouvert.

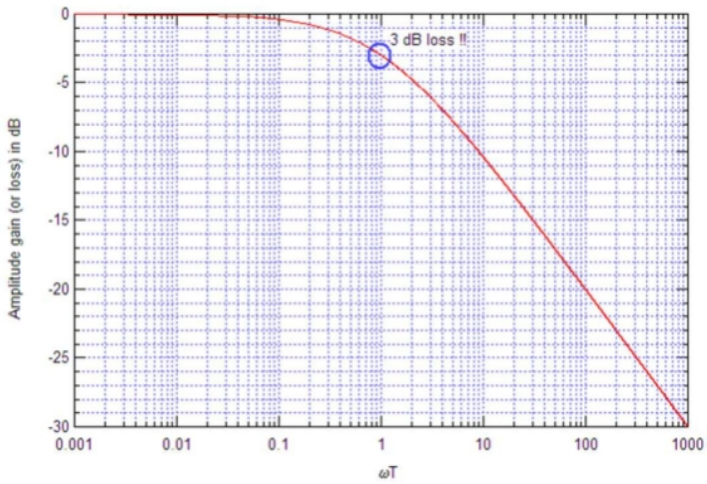


Figure 2.11

- L'impédance équivalente Z_{Th} est l'impédance que le circuit entre les bornes A et B aurait si toutes les sources de tension idéales du circuit étaient remplacées par un court-circuit et toutes les sources de courant idéales étaient remplacées par un circuit ouvert.

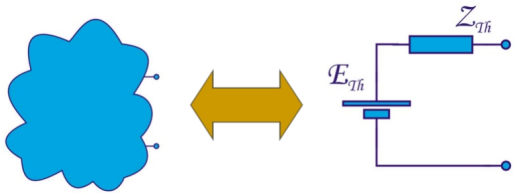


Figure 2.12: Thévenin equivalent

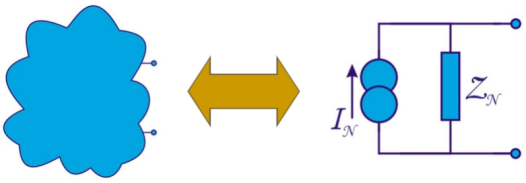


Figure 2.13: Norton equivalent

2.4.2 Le théorème de Norton

Le théorème de Norton nous indique comment remplacer un réseau linéaire par une source de courant I_N en parallèle avec une impédance Z_N, comme illustré dans la figure 2.13. C'est le dual du théorème de Thévenin.

- Le courant de Norton I_N est calculé comme le courant qui circule aux bornes d'un court-circuit (résistance nulle entre les bornes de sortie).

- L'impédance Z_N est trouvée en calculant la tension de sortie produite sans résistance connectée aux bornes; de manière équivalente, c'est la résistance entre les bornes avec toutes les sources de tension (indépendantes) court-circuitées et toutes les sources de courant (indépendantes) circuit ouvert. Cela équivaut à calculer la résistance de Thévenin : $Z_{Th} = Z_N$.

De plus, la tension de sortie générée par la source de Norton avec les bornes de sortie ouvertes est égale à la tension de Thévenin : $E_{Th} = Z_N; I_N$.

2.4.3 Deux Ports

Certains circuits ont deux noeuds d'accès: une entrée et une sortie. Nous supposons que le port est unilatéral: les signaux transitent de l'entrée à la sortie et ne reviennent pas. Nous modélisons l'entrée comme une impédance Z_i et la sortie comme un circuit équivalent Thévenin ou Norton - voir la figure 2.14. Notez comment les sources de courant et de tension dépendent toutes deux de la tension d'entrée. Les différents éléments ont des noms:

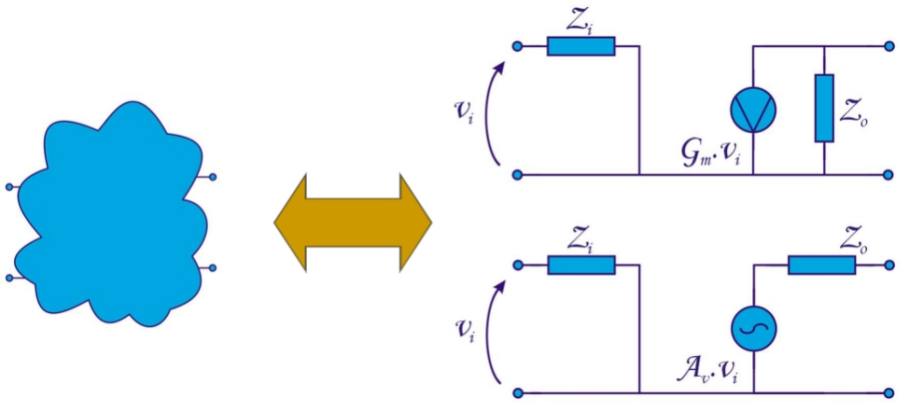


Figure 2.14

- Z_i: l'impédance d'entrée,
- Z_o: l'impédance de sortie,
- G_m: la transconductance,
- A_v: le gain de tension.

Ils peuvent être calculés de la manière suivante:

- Z_i: appliquer une tension d'entrée V_i, calculer le courant d'entrée I_i. Ensuite, $Z_i = \frac{V_i}{I_i}$;
- Z_o: éliminer la transconductance en reliant l'entrée à la masse: $V_i = 0$, appliquer une tension V_o à la sortie et calculer le courant de sortie I_o. Ensuite, $Z_o = \frac{V_o}{I_o}$.
- G_m: court-circuiter la sortie de sorte qu'aucun courant ne passe à travers Z_o. Appliquer une tension d'entrée V_i et calculer le courant de sortie I_o. Ensuite, $G_m = \frac{I_o}{V_i}$.
- A_v: garder la sortie ouverte. Appliquer une tension d'entrée V_i et calculer la tension de sortie V_o. Ensuite, $A_v = \frac{V_o}{V_i}$.

2.4.4 Circuits en cascade

Lorsque plusieurs circuits sont connectés en cascade, on parle de circuits en cascade. Ils se composent généralement de :

- Une source, telle qu'une antenne, un générateur de signal ou un microphone, que l'on modélise comme une source de courant i_s ou de tension v_s, en parallèle ou en série avec une impédance de sortie Z_s.

- Un étage intermédiaire tel qu'un amplificateur ou un filtre qui transforme réellement le signal. Cet étage a une impédance d'entrée Z_i, une impédance de sortie Z_o, un gain de tension A_v ou une transconductance G_m.
- Une charge, qui peut être un PC, un oscilloscope, un haut-parleur, ... Nous modélisons cela comme une impédance de charge Z_l.

Cette configuration est illustrée dans la figure **??**.
Si nous ne faisons pas attention, chaque étage influencera l'étage précédent ou suivant. Nous pouvons calculer la relation entre le courant de source i_s et la tension de charge v_l dans la figure **??** :

- La tension v_i aux bornes de Z_i est :

$$v_i = i_s \ (Z_s||Z_i) = \frac{Z_s Z_i}{Zs + Z_i} i_i$$

  et pas simplement $i_s; Z_s$. L'étage intermédiaire charge l'étage source.

- La tension de charge v_l aux bornes de Z_l est :

$$v_l = -G_m(Z_o||Z_l)v_i = -G_m \frac{Z_o Z_l}{Z_o + Z_l} v_i$$

  et pas simplement $-G_m Z_l v_i$. L'impédance de sortie de l'étage intermédiaire s'ajoute à la charge.

- Cela signifie que la tension du signal disponible en sortie est :

$$v_l = -G_m Z_o \frac{Z_l}{Z_o + Z_l} \frac{Z_i}{Zs + Z_i} Z_s i_i = \frac{Z_l}{Z_o + Z_l} \frac{Z_i}{Zs + Z_i} A_v v_s$$

Idéalement, cela devrait être $v_l = -A_v v_s$. À partir de cela, nous pouvons conclure que $Z_i \to \infty$ et que $Z_o \to 0$ pour se rapprocher du comportement idéal. Nous devons concevoir pour avoir une impédance d'entrée élevée et une impédance de sortie faible.

Part I

Components

Chapter 3

Théorie des Semi-conducteurs et des Solides

Dans ce chapitre, nous aborderons le concept de semi-conducteur. Nous décrirons le type de matériau que cela représente, basé sur le concept de bandes d'énergie. Sous forme solide, un semi-conducteur comme le silicium forme un cristal, donc une attention particulière est portée au comportement des électrons dans un cristal. Le nombre de porteurs de charge dans un semi-conducteur (qui peuvent être des électrons, comme dans un métal, ou des trous, qui n'existent pas dans un métal) est calculé. Pour augmenter le nombre de porteurs, nous dopons le semi-conducteur avec des impuretés. Ensuite, nous considérons les phénomènes de transport les plus importants dans les semi-conducteurs, à savoir la dérive et la diffusion. Enfin, nous discutons de la génération et de la recombinaison de paires électron-trou et établissons les équations de continuité.

3.1 Les Semi-conducteurs

Nous appelons semi-conducteurs les éléments de la quatrième colonne de la table périodique. Des exemples sont le carbone (C), le silicium (Si) et le germanium (Ge). Ces éléments ont 4 électrons sur leur couche externe et ont tendance à former 4 liaisons covalentes avec les atomes voisins pour obtenir une structure octet stable. En se liant, ils forment un motif régulier appelé cristal.
Le silicium, l'atome de semi-conducteur le plus couramment utilisé en électronique, a une couche externe ($n = 3$) avec un orbitale $3s^2$ complètement rempli et un orbitale $3p^2$ partiellement rempli. Pour former des liaisons, les orbitales de la couche externe interagissent et subissent une hybridation sp^3. La conséquence est que les 4 liaisons covalentes sont uniformément espacées dans l'espace, formant des liaisons d'environ 109° les unes avec les autres. La partie gauche de la figure 3.1 montre une seule cellule de la structure cristalline du silicium. Les points bleus représentent un seul atome de silicium, avec des liaisons covalentes à 4 atomes voisins. La figure de droite montre une représentation simplifiée en 2 dimensions du même cristal.
Parfois, un élément composé peut également être un semi-conducteur. L'arséniure de gallium (GaAs) est un composé composé d'atomes de gallium et d'arsenic. C'est un semi-conducteur composé III-V, ce qui signifie qu'il est composé d'éléments des groupes III et V de la table périodique.

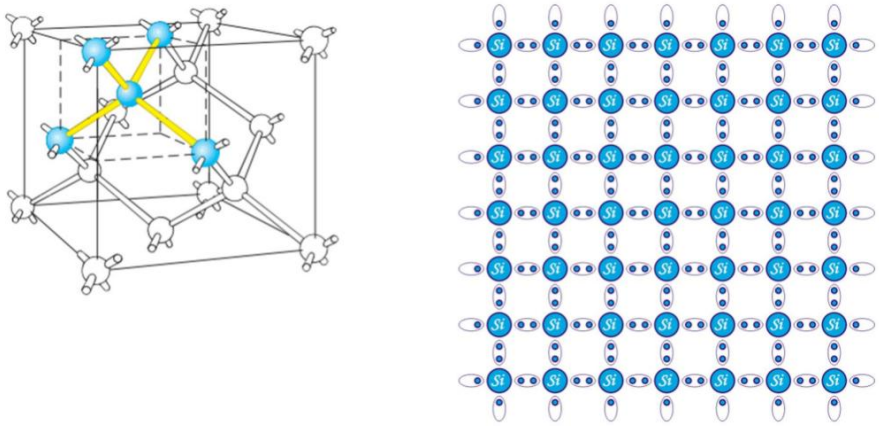


Figure 3.1: Silicon crystal (left) - schematic representation (right)

3.1.1 Structure de bande

Lorsque les atomes sont très éloignés, ils ne s'influencent pas mutuellement et nous pouvons trouver leur fonction d'onde $\Psi(x,t)$ en résolvant l'équation de Schrödinger pour un seul électron en orbite autour d'un noyau fixe. Cela résulte dans le spectre d'énergie discret bien connu associé à chaque fonction d'onde. Cependant, lorsque les atomes se rapprochent, leurs électrons dans la couche externe commencent à interagir les uns avec les autres. Selon le principe d'exclusion de Pauli, ils ne peuvent plus occuper les mêmes niveaux d'énergie.
Supposons que N atomes de Si sont initialement très éloignés, et nous les rapprochons avec pour objectif de former un cristal. Nous avons donc $2N$ électrons au niveau d'énergie de l'orbitale $3s$, et $2N$ électrons dans l'orbitale $3p$, où il y a $6N$ niveaux disponibles. Lorsque les atomes sont suffisamment proches, ils interagissent et les niveaux d'énergie doivent légèrement descendre ou monter pour être conformes au principe d'exclusion de Pauli. Les niveaux sont toujours discrets, mais étant donné que N, le nombre d'atomes dans le cristal, est un nombre énorme ($N \approx 10^{20}$ atomes par cm^3), nous pouvons considérer la bande d'énergie résultante comme un spectre continu (voir la figure 3.2). À partir d'un certain espacement de réseau, les bandes $3s$ et $3p$ fusionnent pour former une seule bande. Lorsque les atomes se rapprochent suffisamment pour former un cristal comme dans la figure 3.1, nous observons que les bandes se séparent à nouveau. À la distance interatomique dans le réseau cristallin ($5.43\mathring{A}$ pour le silicium), il y a une *bande de valence* avec des énergies jusqu'à une énergie E_V et une *bande de conduction* avec des énergies supérieures à $E_C > E_V$. Entre les bandes se trouve une plage interdite où aucun électron ne peut exister. Cette plage est appelée *gap* de bande et la différence entre E_C et E_V est l'*énergie de gap* E_g.

3.2 Électrons et trous

À une température de 0 K, tous les $4N$ électrons se trouvent dans la bande de valence. Cela signifie qu'ils font tous partie d'une liaison covalente entre deux atomes de silicium. Cependant, à mesure que la température augmente, les électrons obtiennent plus d'énergie thermique et certains d'entre eux peuvent rompre la liaison covalente et devenir des électrons libres qui peuvent se déplacer à travers le cristal. Ils ont acquis suffisamment d'énergie pour

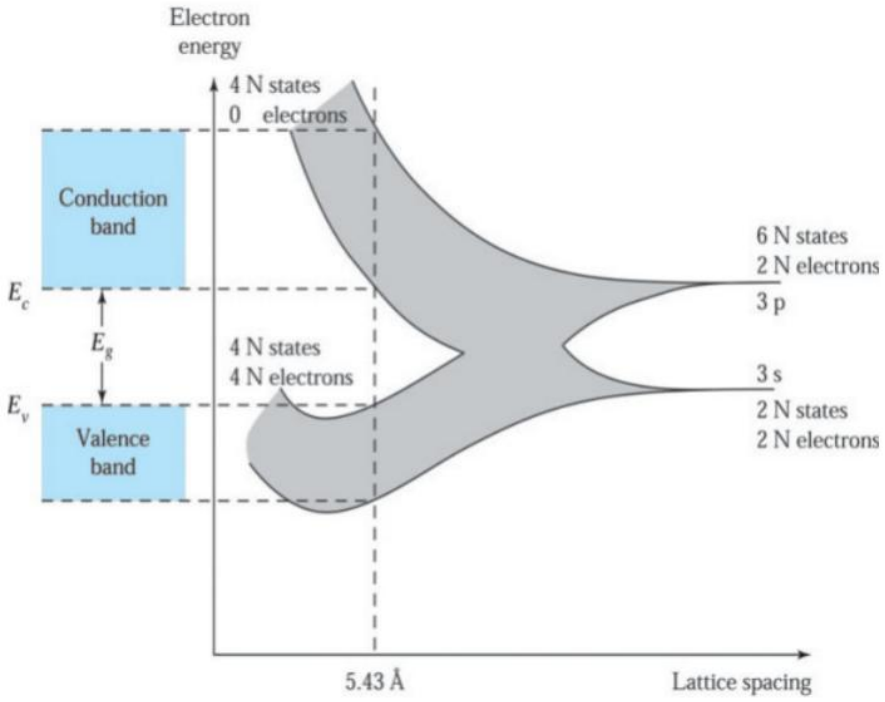


Figure 3.2: Band structure formation

passer de la bande de valence à la bande de conduction. Lorsqu'une liaison covalente est rompue, l'électron laisse derrière lui une position vide. Cette position libre peut être occupée par un autre électron provenant d'une autre liaison électronique. Ce mécanisme est illustré sur la figure **??**. Une position vide dans une liaison est appelée un *trou*. Puisque les électrons sautant d'un trou à l'autre font bouger le trou dans le cristal, nous pouvons considérer un trou comme une autre particule en mouvement, avec une charge positive $+q$.
Il existe également d'autres mécanismes en plus de l'agitation thermique qui peuvent fournir suffisamment d'énergie à un électron pour qu'il saute de la bande de valence à la bande de conduction, comme un photon qui frappe avec une énergie E suffisamment élevée (donc une fréquence ν, puisque $E = \hbar\nu$). Le fond de la bande de conduction E_C correspond à l'énergie potentielle d'un électron, tout comme le sommet de la bande de valence E_V correspond à l'énergie potentielle d'un trou.
Lorsque l'on considère différentes directions dans le réseau cristallin, la structure de bande diffère car la distance interatomique dépend de la direction. La direction de propagation d'une fonction d'onde est déterminée par son vecteur d'onde $\vec{k}$, et k est directement lié à l'impulsion p par la relation :

$$\vec{p} = \hbar\vec{k}$$

C'est pourquoi ces *diagrammes de bande* sont généralement tracés en fonction de l'énergie E en fonction de l'impulsion p. La figure 3.4 montre les diagrammes de bande pour le silicium (à gauche) et le GaAs (à droite). Ces figures montrent que pour certains semi-conducteurs, comme le GaAs, le minimum de la bande de conduction se situe dans la même direction que le maximum de la bande de valence. Ces semi-conducteurs sont appelés semi-conducteurs à *bande interdite directe*. D'autres, comme le silicium, ont des directions différentes pour ces

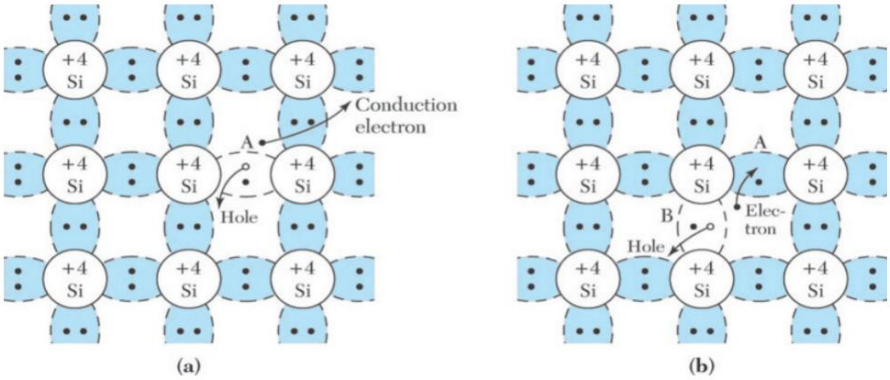


Figure 3.3: Creation of electron hole pair by breaking of covalent bonds

deux points et sont donc appelés semi-conducteurs à *bande interdite indirecte*.

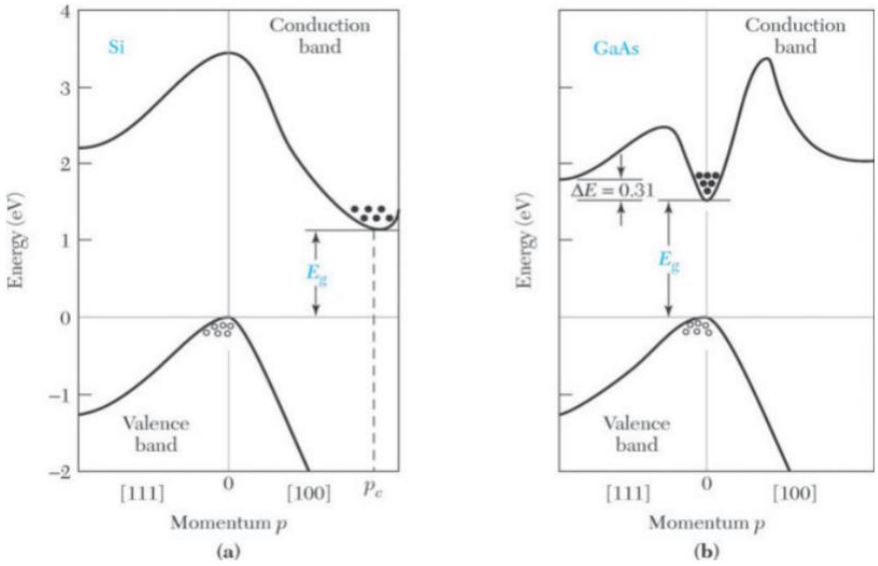


Figure 3.4: Band diagrams: (a) Si and (b) GaAs

3.2.1 Masse effective

Un électron libre - non entravé par d'autres particules et leurs fonctions potentielles - a une masse $m_0 = 9,11\ 10^{-31}$ kg. Cependant, dans un cristal, les électrons de la bande de conduction sont influencés par les noyaux et les autres électrons. Pour déterminer leur comportement, nous devrions résoudre l'équation de Schrödinger pour toutes ces particules. À la place, nous introduisons le concept de masse effective, qui peut être considérée comme la masse apparente d'une particule (électron ou trou) dans un cristal.
Chaque fonction d'onde a une vitesse de groupe $v_g = \frac{d\omega}{dk}$ et puisque $E = \hbar\omega$, nous avons $v_g = \frac{1}{\hbar}\frac{dE}{dk}$. Mais comme l'augmentation d'énergie dE de la particule peut être considérée comme le résultat du travail dW effectué par une force F via la relation $dW = Fdl = Fv_g dt$ où dl est le déplacement en temps dt, nous pouvons simplifier cette relation en $\frac{dk}{dt} = \frac{F}{\hbar}$.

Si une particule subit une variation de vitesse de groupe - une accélération a - nous pouvons affirmer que:

$$\begin{aligned} a &= \frac{dv_g}{dt} = \frac{1}{\hbar}\frac{d}{dt}\frac{dW}{dk} \\ &= \frac{1}{\hbar}\frac{d^2W}{dk^2}\frac{dk}{dt} \\ &= \frac{1}{\hbar^2}\frac{d^2W}{dk^2}F \end{aligned} \tag{3.1}$$

Cela donne une relation entre la force appliquée F et l'accélération résultante a. Ainsi, selon la deuxième loi de Newton $F = ma$, nous pouvons interpréter la constante de proportionnalité comme la masse effective de la particule:

$$m^* = \hbar^2 \Big[\frac{d^2W}{dk^2}\Big]^{-1}$$

La masse effective d'un électron ou d'un trou est donc déterminée par la courbure locale de la relation $E - k$ dans le diagramme de bande. À partir de la figure 3.4, nous voyons que pour le silicium, la courbure de la bande de conduction est plus élevée que celle de la bande de valence. Par conséquent, la masse effective m_e^* d'un électron sera plus petite que celle d'un trou m_h^*.

3.2.2 Nombre de porteurs

Nous pouvons calculer la concentration d'électrons et de trous dans le silicium pur en utilisant la statistique quantique. La distribution des fermions (comme les électrons) est déterminée par la distribution de Fermi-Dirac :

$$F(E) = \frac{1}{1 - e^{(E-E_F)/kT}}$$

qui donne la probabilité qu'à la température T, un électron occupe le niveau d'énergie E. E_F est le niveau de Fermi, un niveau de référence où la probabilité d'occupation est exactement de moitié, et k est la constante de Boltzmann (à ne pas confondre avec le nombre d'onde k). Nous considérons également le nombre d'états autorisés à un niveau d'énergie E. Nous savons déjà que entre E_V et E_C, aucun état n'est disponible car c'est la région de la bande interdite. Pour respectivement les électrons dans la bande de conduction et les trous dans la bande de valence, nous avons que :

$$\begin{aligned} N_C(E) &= \frac{4\pi}{\hbar}(2m_e^*)^{3/2}(E - E_C)^{1/2} \\ N_V(E) &= \frac{4\pi}{\hbar}(2m_h^*)^{3/2}(E_V - E)^{1/2} \end{aligned} \tag{3.2}$$

où $N_C(E)$ et $N_V(E)$ représentent la densité d'états par unité de volume dans la bande de conduction et la bande de valence. Pour calculer la densité de porteurs, nous multiplions la densité d'états par la probabilité qu'un état soit occupé, et intégrons sur les niveaux d'énergie pertinents. Nous notons n pour le nombre d'électrons dans la bande de conduction, et p pour

le nombre de trous dans la bande de valence :

$$\begin{aligned} n &= \int_{E_C}^{\infty} N_C(E) F_e(E) dE \\ p &= \int_{-\infty}^{E_V} N_V(E) F_h(E) dE \end{aligned} \tag{3.3}$$

Puisque E_C et E_V sont tous deux suffisamment éloignés du niveau de Fermi, nous pouvons simplifier la distribution de Fermi-Dirac et obtenir une distribution de Boltzmann standard :

$$\begin{aligned} F_e(E) &\approx e^{-(E-E_F)/kT} \text{ if } E - E_F \gg kT \\ F_h(E) &\approx e^{(E-E_F)/kT} \text{ if } E - E_F \ll kT \end{aligned} \tag{3.4}$$

En substituant 3.4 dans 3.3, on obtient :

$$\begin{aligned} n &= N_C \, e^{-(E_C-E_F)/kT} \text{ with } N_C = \ldots \\ p &= N_V \, e^{-(E_F-E_V)/kT} \text{ with } N_V = \ldots \end{aligned} \tag{3.5}$$

La figure 3.5 représente le schéma de bande, la densité d'états $N(E)$, la distribution des porteurs $F(E)$ et les résultats finaux.

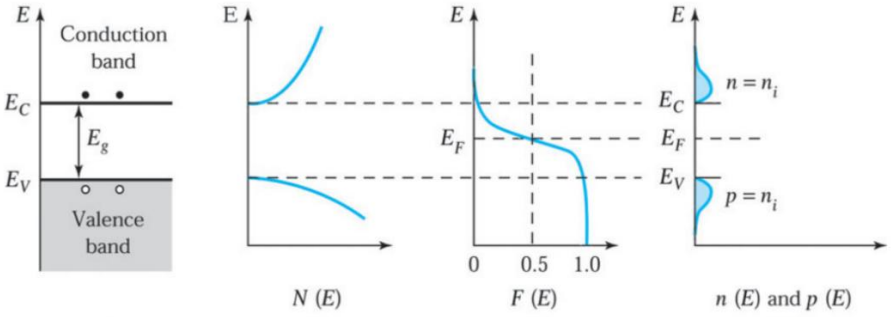


Figure 3.5: Carrier density

Dans le silicium pur, chaque électron dans la bande de conduction donne naissance à un trou dans la bande de valence - d'où $n = p$ et E_F se situe presque à mi-chemin entre E_V et E_C (la petite différence est due à une différence de masse effective entre les trous et les électrons). On dit que $n = p = n_i$, la densité de porteurs intrinsèques. En général, à température ambiante, $n \approx 10^{10}/cm^3$. Si nous multiplions les équations pour n et p, le niveau de Fermi disparaît :

$$np = n_i^2 = N_C \, e^{-(E_C-E_F)/kT} \, N_V \, e^{-(E_F-E_V)/kT} = N_C N_V \, e^{-E_g/kT} \tag{3.6}$$

3.3 Impuretés donneuses et accepteuses

La densité de porteurs intrinsèques n_i est assez faible, et le silicium pur est un mauvais conducteur. Cependant, nous pouvons augmenter le nombre de porteurs (soit des trous, soit des électrons) avec un processus appelé *dopage*. Le dopage consiste à remplacer certains des atomes de silicium par des atomes du groupe III (atomes accepteurs) ou du groupe V (atomes donneurs) du tableau périodique. Si les atomes de remplacement sont des atomes

donneurs (comme l'arsenic ou le phosphore), ils se lient aux atomes de silicium voisins mais ont toujours un électron supplémentaire. Cet électron n'est que faiblement lié au noyau et peut facilement passer à la bande de conduction - sans créer de trou. De même, les atomes du groupe III, comme le bore, ont un électron en moins pour créer 4 liaisons covalentes. Cependant, s'ils peuvent arracher un électron à une autre liaison, ils peuvent l'utiliser pour obtenir une structure octet et en même temps ont créé un trou. Les deux processus sont schématiquement représentés sur la figure 3.6. Le dopage élimine également la dépendance de la concentration de porteurs majoritaires par rapport à la température.

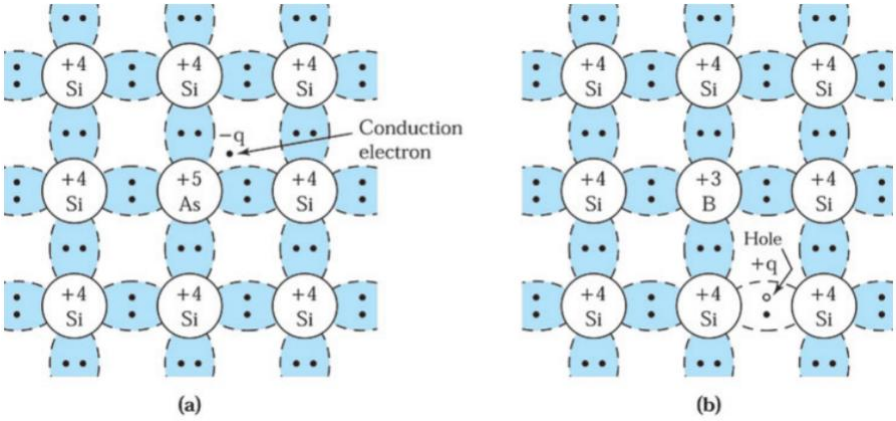


Figure 3.6: Doping with (a) donor atoms and (b) acceptor atoms

Conventionnellement, nous utilisons N_d pour la concentration de donneurs et N_a pour la densité d'accepteurs. Ils ont des valeurs typiques de $10^{16}/cm^3$. Cela est beaucoup plus faible que le nombre d'atomes ($\sim 10^{20}/cm^3$), mais beaucoup plus élevé que la concentration intrinsèque de porteurs n_i ($\sim 10^{10}/cm^3$). Nous parlerons de semi-conducteurs de type n ou p lorsque nous parlons de dopage avec des donneurs ou des accepteurs, respectivement.
L'expression 3.5 reste valable, également dans un semi-conducteur dopé. Par conséquent, les porteurs de charge majoritaires (électrons dans un semi-conducteur de type n, trous dans un semi-conducteur de type p) sont largement plus nombreux que les porteurs de charge minoritaires (trous dans un semi-conducteur de type n, électrons dans un semi-conducteur de type p). Dans le cas d'un semi-conducteur de type n, par exemple, nous supposons que tous les atomes donneurs perdent leur électron supplémentaire et deviennent ionisés. Ainsi, $n_n = N_d \approx 10^{16}/cm^3$. En conséquence, $p_n = n_i^2/n_n \approx 10^4/cm^3$. Notez le sous-script n pour indiquer qu'il s'agit d'un semi-conducteur de type n.
Le niveau de Fermi pour les semi-conducteurs dopés (ou *extrinsèques*) ne se situe plus à mi-chemin entre la bande de valence et la bande de conduction. Puisque:

$$n_n = N_d = N_C e^{-(E_C - E_{Fn})/kT} \Rightarrow E_{Fn} = E_C - kT \ln \frac{N_d}{N_C}$$

et nous savons également que $E_i = E_C - kT \ln \frac{n_i}{N_C}$ pour un semi-conducteur intrinsèque. Ainsi:

$$E_{Fn} = E_i + kT \ln \frac{N_d}{n_i}$$

and

$$E_{Fp} = E_i - kT \ln \frac{N_a}{n_i}$$

Cela signifie que dans un semi-conducteur de type n, le niveau de Fermi se situe au-dessus du niveau de Fermi intrinsèque, tandis que dans un semi-conducteur de type p, il se situe en dessous. De plus, le niveau de Fermi se rapproche de la bande de conduction (de valence) si N_d (N_a) est plus élevé. La figure 3.7 montre comment la distribution de charges change dans un semi-conducteur de type n, en raison d'un déplacement du niveau de Fermi E_{Fn}. Notez également le surplus d'électrons par rapport aux trous, puisque presque tous les électrons de la bande de conduction proviennent de l'ionisation des atomes donneurs.

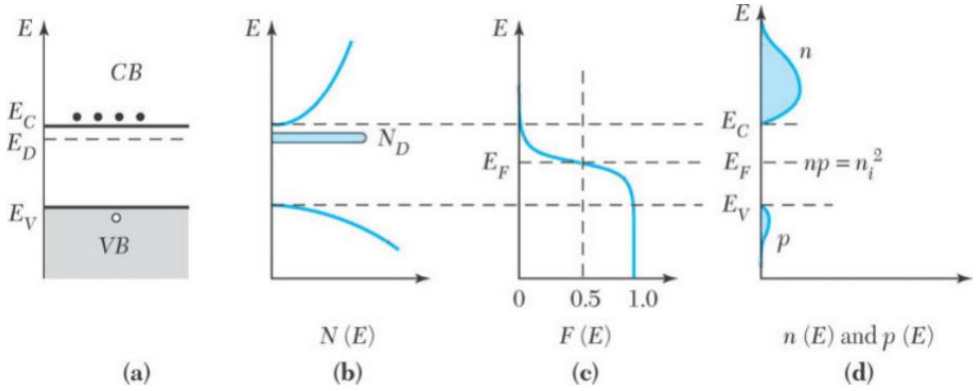


Figure 3.7: Carrier density due to donor doping in n-type semiconductor

Si nous transformons les énergies en potentiels via la relation $E = -qV$ et définissons V comme la différence de potentiel entre le niveau de Fermi intrinsèque et réel : $V = V_{Fi} - V_F$, nous pouvons réécrire ces équations et obtenir les *équations de Boltzmann*:

$$\begin{aligned} n &= n_i\, e^{qV/kT} = n_i e^{V/v_{th}} \\ p &= n_i\, e^{-qV/kT} = n_i e^{-V/v_{th}} \end{aligned} \tag{3.7}$$

avec $v_{th} = kT/q$ la tension thermique (≈ 26 mV à $T = 300$ K). Ces équations ne sont valables qu'à l'équilibre thermique, c'est-à-dire lorsqu'il n'y a pas d'écoulement de charges.

3.3.1 Transport de Porteurs

Plusieurs mécanismes de transport de porteurs existent dans un semi-conducteur. Nous discutons ici les deux plus courants : le courant de dérive, dû à la présence d'un champ électrique, et le courant de diffusion, dû à un gradient de concentration.

3.3.2 Courant de Dérive

En l'absence d'un champ électrique, les charges se déplacent de manière aléatoire en raison de leur énergie thermique. Cependant, puisque le mouvement est aléatoire, il n'existe aucune direction préférentielle de déplacement. Cela change lorsque un champ électrique externe est appliqué. En présence d'un champ électrique $\mathcal{E}$, un électron avec une charge $-q$ est accéléré par une force $F_1 = -q\mathcal{E}$. Cependant, en même temps, la charge interagit avec le réseau cristallin et est ralentie par des collisions avec les atomes et les impuretés. Cette force d'amortissement F_2 est proportionnelle à la vitesse de la particule : $F_2 = -\alpha v_d$, avec α le facteur d'amortissement. Ainsi :

$$m_e^* \frac{dv_d}{dt} = -q\mathcal{E} - \alpha v_d$$

La particule atteindra une vitesse d'équilibre v_{d0} lorsque $\frac{dv_d}{dt} = 0$, ainsi $v_{d0} = \frac{-qE}{\alpha}$. Nous pouvons résoudre l'équation du premier ordre pour une particule qui est initialement à une vitesse $v_d(0) = 0$ et obtenir $v_d(t) = v_{d0}e^{-t/\tau_e}$ où $\tau_e = \frac{m_e^*}{\alpha}$ est un temps caractéristique appelé *temps de relaxation*. Il est proportionnel au temps nécessaire pour atteindre v_{d0} et est de l'ordre de 1 ps. D'un point de vue macroscopique, pour calculer le courant électronique total, nous devons faire la moyenne sur tous les électrons disponibles pour la conduction, c'est-à-dire les électrons présents dans la bande de conduction. Après quelques calculs, nous obtenons pour la densité de courant électronique $J_n = q\mu_n n\mathcal{E}$ avec la *mobilité des électrons* $\mu_n = \frac{q\tau_e}{m_e^*}$. Une expression similaire peut être trouvée pour la densité de courant de trous J_p. Le courant total est la somme des deux:

$$J = J_n + J_p = q(\mu_n n + \mu_p p)\mathcal{E} = \sigma\mathcal{E} \tag{3.8}$$

Ceci est la formulation de la loi d'Ohm pour un semi-conducteur.

3.3.3 Structure de bande sous polarisation

Nous considérons la conduction dans un matériau semi-conducteur homogène due à un champ électrique. La figure 3.8 montre un semi-conducteur de type n et son diagramme de bande à l'équilibre thermique (à gauche) et le diagramme de bande lorsqu'une tension de polarisation positive est appliquée à la borne de droite (à droite). Lorsqu'un champ électrique E est appliqué à un semi-conducteur, chaque électron subit une force $-q\mathcal{E}$ du champ. La force est égale au gradient négatif de l'énergie potentielle :

$$-q\mathcal{E} = \frac{dE_C}{dx}$$

Puisque nous nous intéressons au gradient de l'énergie potentielle, nous pouvons utiliser n'importe quelle partie du diagramme de bande qui est parallèle à E_C. Il est pratique d'utiliser le niveau de Fermi intrinsèque E_i car nous l'utiliserons lorsque nous considérerons les jonctions p-n. Ainsi:

$$\mathcal{E} = -\frac{1}{q}\frac{dE_C}{dx} = -\frac{1}{q}\frac{dE_i}{dx}$$

Nous pouvons définir une quantité associée V comme le potentiel électrostatique dont le gradient négatif est égal au champ électrique: $\mathcal{E} = -\frac{dV}{dx}$. En comparant les deux équations, nous obtenons $V = -\frac{E_i}{q}$, ce qui établit une relation entre le potentiel électrostatique et l'énergie potentielle d'un électron. Pour le semi-conducteur homogène illustré à la figure 3.8, l'énergie potentielle et E_i diminuent linéairement avec la distance ; ainsi, le champ électrique est constant dans la direction négative de x. Sa magnitude est la tension appliquée divisée par la longueur de l'échantillon.

Les électrons de la bande de conduction se déplacent vers le côté droit. L'énergie cinétique correspond à la distance par rapport au bord de bande (c'est-à-dire E_C pour les électrons). Lorsqu'un électron subit une collision, il perd une partie ou la totalité de son énergie cinétique vers le réseau cristallin et retourne à sa position d'équilibre thermique. C'est l'origine du facteur d'atténuation α de la section précédente. Après que l'électron a perdu une partie ou la totalité de son énergie cinétique, il commencera à nouveau à se déplacer vers la droite et le même processus se répétera plusieurs fois. La conduction par les trous peut être visualisée de manière similaire mais dans la direction opposée.

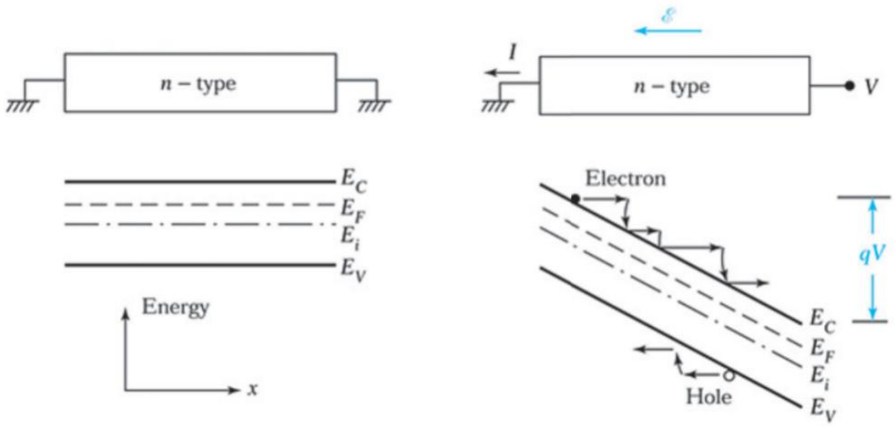


Figure 3.8: Band structure under biasing

3.3.4 Courant de diffusion

Un autre mécanisme de transport est la diffusion où les porteurs se déplacent d'un endroit à un autre en raison de la variation spatiale de la concentration de porteurs. Comme les porteurs se déplacent au hasard en raison de l'agitation thermique, plus de porteurs se déplaceront de la concentration de porteurs plus élevée vers celle de porteurs plus faible que dans l'autre sens, entraînant efficacement un mouvement net d'électrons de gauche à droite comme dans la figure 3.9. Ce courant est décrit par une équation de diffusion standard :

$$J_n = qD_n \frac{dn}{dx}$$

avec D_p le coefficient de diffusion des trous. Une relation equivalente existe pour les trous:

$$J_p = -qD_p \frac{dp}{dx}$$

Cependant, ce déplacement de porteurs induira un champ électrique $\mathcal{E}$ et donc également

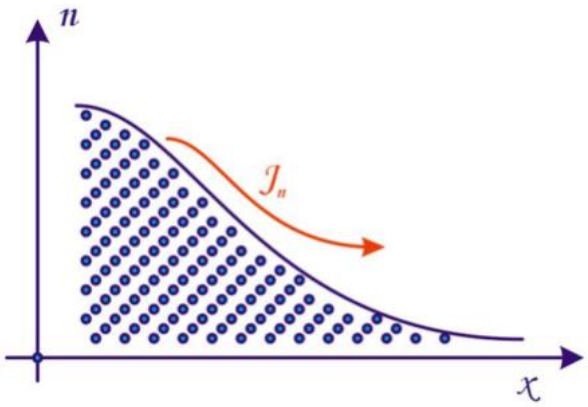


Figure 3.9: Courant J_n dû à la diffusion

un courant de dérive $J = q\mu_n n\mathcal{E}$. Après un certain temps, les deux courants seront égaux et

opposés et un équilibre dynamique sera atteint :

$$J_n = q\mu_n n\mathcal{E} + qD_n\frac{dn}{dx} = 0$$

En utilisant les équations de Boltzmann (3.7), on peut déduire une relation entre D_n et μ_n :

$$D_n = \frac{kT}{q}\mu_n$$

et

$$D_p = \frac{kT}{q}\mu_p$$

Ces relations sont connues sous le nom d'équations d'Einstein.

3.4 Génération et Recombinaison

En équilibre thermique, la relation $pn = n_i^2$ est valide. Il s'agit d'un équilibre dynamique : la génération thermique de paires électron-trou à un taux G_{th} est contrebalancée par les électrons qui retombent de la bande de conduction à la bande de valence (c'est-à-dire un électron libre qui se combine avec un trou pour former une liaison covalente). Ce processus s'appelle la recombinaison.
Si des porteurs en excès sont introduits, nous ne sommes plus en équilibre et $pn > n_i^2$. La création de porteurs en excès est appelée *injection de porteurs* et peut être réalisée par excitation optique ou en polarisant directement une jonction pn (voir le chapitre 4). Le mécanisme qui rétablit l'équilibre est la recombinaison des porteurs minoritaires injectés avec les porteurs majoritaires présents dans le semi-conducteur. La génération et la recombinaison de porteurs thermiques et en excès sont représentées dans la figure **??**. Nous supposons que

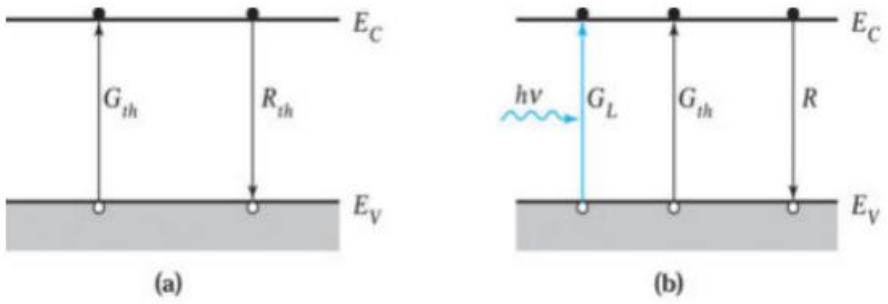


Figure 3.10: Generation and recombination under (a) equilibrium and (b) optical excitation

les porteurs en excès conduisent à une recombinaison à une vitesse $R(n, p, T)$. À l'équilibre, $G_{th}(n_0, p_0, T) = R(n_0, p_0, T) = R_{th}$. Lorsque nous créons des porteurs en excès, nous pouvons écrire que $n = n_0 + \delta n$ et $p = p_0 + \delta p$, et nous supposerons que l'injection est faible par rapport aux conditions d'équilibre. Cela signifie que $\delta n \ll n_0$ et $\delta p \ll p_0$. Dans ces conditions, nous pouvons approximer la vitesse de recombinaison par les termes du premier ordre:

$$\begin{aligned} R(n,p,T) &= R(n_0,p_0,T) + (n-n_0)\frac{\partial R}{\partial n} + (p-p_0)\frac{\partial R}{\partial p} \\ &= G_{th} + \frac{(n-n_0)}{\tau_n} + \frac{(p-p_0)}{\tau_p} \end{aligned} \tag{3.9}$$

Dans un matériau de type n, le taux de recombinaison est déterminé par les trous en excès - car les trous sont beaucoup plus susceptibles de trouver un candidat pour la recombinaison (c'est-à-dire un électron) - que par les électrons en excès. Le taux auquel les porteurs recombinent est donc déterminé par le nombre de porteurs minoritaires :

- En type n : $R - G_{th} = \frac{(p-p_0)}{\tau_p}$
- En type p : $R - G_{th} = \frac{(n-n_0)}{\tau_n}$

La durée de vie des porteurs minoritaires en excès τ_p et τ_n est beaucoup plus grande que le temps de relaxation τ_e.

3.5 Les équations de continuité

Le changement de densité de porteurs dans un volume V à l'intérieur d'un semi-conducteur peut être dû à trois causes :

1. génération de porteurs,
2. recombinaison,
3. flux de porteurs entrant ou sortant du volume à travers la surface environnante S

Supposons que nous voulons étudier le changement de la densité d'électrons dans un matériau de type p. Nous pouvons écrire:

$$\begin{aligned} -q \iiint_V \frac{\partial n}{\partial t}\, dV &= -q \iiint_V (G - R)\, dV - \oiint_S \vec{J_n}\vec{n}\, dS \\ &= -q \iiint_V (G - R)\, dV - \iiint_V \nabla \vec{J_n}\, dV \end{aligned} \tag{3.10}$$

où nous avons utilisé le théorème de la divergence. Comme cela est valable pour n'importe quel volume, les termes dans les intégrales doivent être égaux:

$$\begin{aligned} -q\frac{\partial n}{\partial t} &= -q(G - R) - \nabla \vec{J_n} \\ &= -q(G_{th} + g - R) - \nabla \vec{J_n} \\ \Rightarrow \frac{\partial n}{\partial t} &= \Big(\frac{n - n_0}{\tau_n}\Big) + \frac{1}{q}\nabla \vec{J_n} + g \end{aligned} \tag{3.11}$$

où nous remplaçons $G_{th} - R$ par $\frac{(n-n_0)}{\tau_n}$ puisque nous sommes dans un matériau de type p. En une dimension, cela devient :

$$\frac{\partial n}{\partial t} = \Big(\frac{n - n_0}{\tau_n}\Big) + \frac{1}{q}\frac{\partial J_n}{\partial x} + g \tag{3.12}$$

où J_n a en général une contribution de dérive et de diffusion : $J_n = q\mu_n n\mathcal{E} + qD_n\frac{dn}{dx}$. En substituant cela dans 3.12 - avec à la fois n et $\mathcal{E}$ dépendant de x - nous obtenons:

$$\frac{\partial n}{\partial t} = \Big(\frac{n - n_0}{\tau_n}\Big) - n\mu_n\frac{\partial \mathcal{E}}{\partial x} + \mu_n\mathcal{E}\frac{\partial n}{\partial x} + D_n\frac{\partial^2 n}{\partial x^2} + g \tag{3.13}$$

Cette équation est appelée l'*équation de continuité* pour les porteurs de type n dans un matériau de type p. Des expressions similaires peuvent être trouvées pour les autres cas.

Chapter 4

The pn-junction

In this section we demonstrate that when a junction is formed between a sample of p-type and one of n-type semiconductor, this combination possesses the properties of a rectifier. This two-terminal device is called a *junction diode*. The physical structure and symbol of a pn-junction is shown in figure 4.1

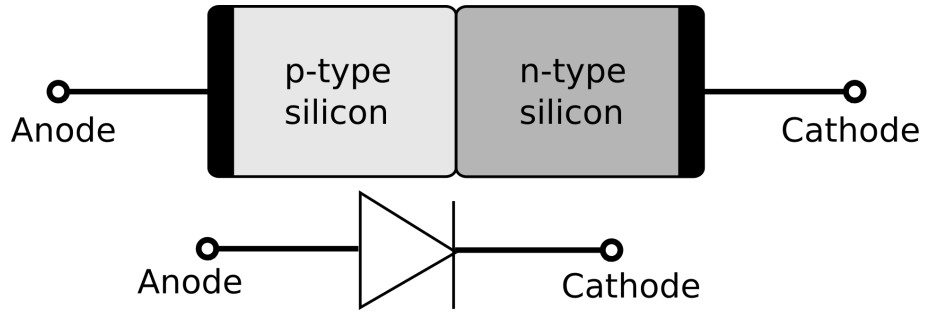


Figure 4.1: Structure (above) and symbol (below) of a pn-junction

We will discuss the junction in equilibrium and under forward and reverse bias. The I-V characteristic will be derived and the concepts of avalanche breakdown and depletion capacitance will be discussed.

4.1 The pn-junction in equilibrium

If donor impurities are introduced into one side and acceptors in the other side of a single crystal of a semiconductor, a *pn*-junction is formed[1]. At the interface, there is by construction a concentration gradient and hence electrons and holes will diffuse to either side. Electrons from the n-type will recombine with the majority holes in the p-type and the holes from the p-type will recombine with the majority electrons in the n-type. As a consequence, the doping ions will be exposed and the n-side will have a fixed positive charge density (N_d) while the p-side will have a fixed negative charge density (N_a) (figure 4.2 (a). This charge buildup will lead to an internal electric field pointing from n-type to p-type and it will thus act against any further diffusion current (4.2 (b)). We assume for now that no external bias is applied,

[1]This is called an *abrupt junction*

so that no net current can flow. The zone where the majority carriers have diffused and recombined is the *space-charge region*[2]. To preserve charge neutrality, we require that:

$$N_A\ x_p = N_d\ x_n$$

with x_p and x_n the depth of the space-charge region in p- and n-type (W_{Dp} and W_{Dn} in figure 4.2). As can be deduced from this expression, the space-charge region extends further in the lightly-doped material.
Because $\frac{d\mathcal{E}}{dx} = \rho/\epsilon$, the electrical field that results from this charge buildup is computed as $\mathcal{E}(x) = \int \rho(x)/\epsilon\ dx$ with $\rho(x)$ the local charge profile:

$$\begin{aligned} \rho(x) &= -N_a \text{ if } x < 0 \\ &= \ N_d \text{ if } x > 0 \end{aligned} \tag{4.1}$$

The maximum electric field corresponds to the total charge buildup and is negative since the p-type is placed left: $|\mathcal{E}_m| = \epsilon N_a x_p = \epsilon N_d x_n$. An electric field gives rise to a potential difference since $\mathcal{E} = -\frac{d\psi}{dx}$ as in figure 4.2(c) where the so-called built-in potential ψ_{bi} is equal to the surface under $\mathcal{E}(x)$:

$$\psi_{bi} = \frac{1}{2}(x_n + x_p)|\mathcal{E}_m| = \frac{1}{2}\epsilon(x_n + x_p)N_a x_p$$

This built-in potential is also visible in 4.2 (d) where the energy bands are shown. Due to electric field, the bands will bend (in opposite direction as for the potential, since $E = -qV$) at the junction. This last figure is important and can also be constructed by reasoning with the Fermi levels.
The total width $W = x_n + x_p$ of the space charge region can be computed as function of the doping levels and the built-in voltage. The result is:

$$W = \sqrt{\frac{2\epsilon}{q}\Big(\frac{N_a + N_d}{N_a N_d}\Big)V_{bi}} \tag{4.2}$$

4.1.1 Fermi-levels in equilibrium

Since no net current can flow in the junction with no external bias, the drift current must be equal to the diffusion current. For the holes, this condition gives:

$$\begin{aligned} J_p &= J_{p,drift} + J_{p,diffusion} \\ &= q\mu_p p\mathcal{E} - qD_p\frac{dp}{dx} \\ &= q\mu_p p\Big(\frac{1}{q}\frac{dE_i}{dx}\Big) - kT\mu_p\frac{dp}{dx} = 0 \end{aligned} \tag{4.3}$$

In equilibrium we can apply the Boltzmann equations:

$$p = n_i e^{(E_i - E_F)/kT} \tag{4.4}$$

[2]Also called the depletion region

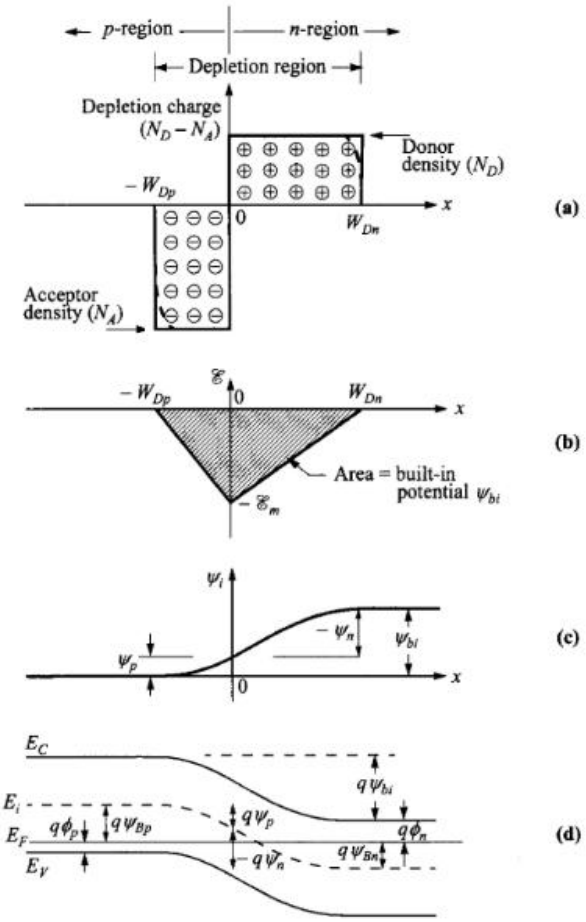


Figure 4.2: pn-junction in thermal equilibrium

and compute $\frac{dp}{dx}$:

$$\frac{dp}{dx} = \frac{p}{kT}\left(\frac{dE_i}{dx} - \frac{dE_F}{dx}\right)$$

Substituting this in equation 4.3 gives:

$$J_p = \mu_p p \frac{dE_F}{dx} = 0$$

or $\frac{dE_F}{dx} = 0$. A similar result is valid for the electrons. **For there to be zero net electron and hole currents, the Fermi level must be constant.**
As can be seen in figure 4.2(d), the Fermi level E_F remains constant along the junction because there is no net current. Since E_F is close to E_V in the p-type and close the E_C in the n-type, the bands must bend like they do in the figure to keep a constant E_F.

4.1.2 The built-in potential

We will now compute $V_{bi} = \psi_n - \psi_p$ (previously denoted as ψ_{bi}). We know that far away from the junction, no net charges are present, so the potential must comply with the Laplace

equation:

$$\frac{d^2\psi}{dx^2} = 0$$

and $N_d - N_a + p - n = 0$ to preserve charge neutrality. In a p-type material, we assume that $N_d = 0$ and $p \gg n$. This results in $p = N_a$. Inserting this in equation 4.4 gives:

$$\psi_p = -\frac{1}{q}(E_i - E_f)|_{x<x_p} = \frac{kT}{q}\ln\left(\frac{N_a}{n_i}\right)$$

Similarly, the electrostatic potential for the n-type material is

$$\psi_n = -\frac{1}{q}(E_i - E_f)|_{x>x_n} = \frac{kT}{q}\ln\left(\frac{N_d}{n_i}\right)$$

The built-in potential V_{bi} is the difference of electrostatic potential between p-side and n-side at thermal equilibrium:

$$V_{bi} = \psi_n - \psi_p = \frac{kT}{q}\ln\left(\frac{N_a N_d}{n_i^2}\right) \tag{4.5}$$

4.2 The pn-junction under bias

4.2.1 Forward bias

Assume we apply a positive voltage V_F to the anode (p-side) while keeping the cathode (n-side) at ground. The effect is that we reduce the barrier of the built-in potential from V_{bi} to $V_{bi} - V_F$, as shown on the left side of figure 4.3. This has a couple of consequences:

- The width of the depletion region will reduce, because equation 4.2 can be rewritten as:

$$W = \sqrt{\frac{2\epsilon}{q}\left(\frac{N_a + N_d}{N_a N_d}\right)(V_{bi} - V_F)} \tag{4.6}$$

- The value of the maximum $\mathcal{E}$ is reduced. This means that there is an disequilibrium between diffusion and drift current, and there will be a net (diffusion) flow of holes from p-type to n-type and a flow of electrons from n-side to p-side. The result is a net current from p-side to n-side.[3]

There is thus a diffusion current of majority carriers through the depletion region to the other side, where they become the minority carriers. Since there are a lot of majority carriers available, this current can become very large when the potential barrier is lowered enough. Also notice that the Fermi levels in figure 4.3 are no longer constant, because there is a current and we are no longer in equilibrium.

4.2.2 Reverse Bias

When we apply a positive voltage V_R to the cathode (n-side) while keeping the anode at ground, we increase the barrier of the built-in potential to $V_{bi} + V_R$. This situation is sketched on the right of figure 4.3. According to equation 4.6, the width of the space-charge region will

[3]Notice that electrons can diffuse from low to higher energy, while they always drift from high to low.

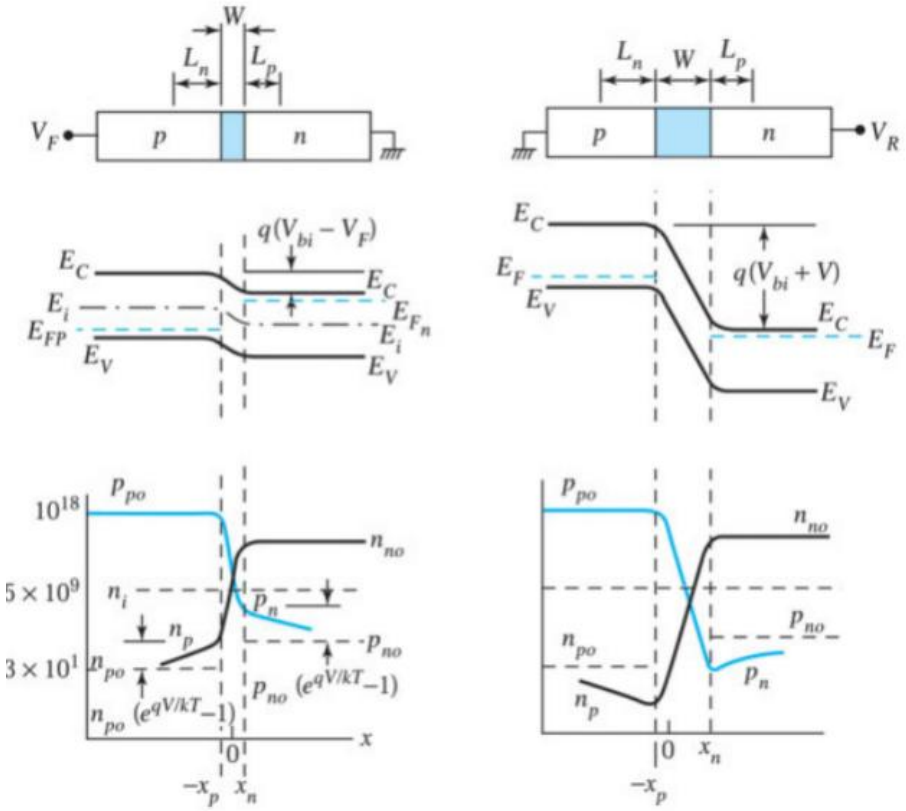


Figure 4.3: pn-junction under forward (left) and reverse (right) bias

increase (because we can replace V_F with $-V_R$). The internal electric field at the junction becomes larger than before, so there can be a net drift current. However, the only available carriers to be swept across the space-charge region by the electric field are the minority carriers on both sides of the junction. Consequently, the resulting current (the *saturation current*) will be very low.

4.2.3 Diode Characteristic

Let's to compute the currents. We'll make the following assumptions:

1. the depletion region has abrupt boundaries and, outside the boundaries, the semiconductor is assumed to be neutral;

2. the carrier densities at the boundaries are related by the electrostatic potential difference across the junction;

3. the injected minority carrier densities are small compared to the majority carrier densities;

4. neither generation nor recombination current exists in the depletion region and the electron and hole currents are constant throughout the depletion region.

At equilibrium, we get $p_{p0} = N_a$ and $n_{n0} = N_d$. Together with the mass-action law $p_{p0}\, n_{n0} = n_i^2$, we can rewrite equation 4.5 as:

$$V_{bi} = \frac{kT}{q} \ln \frac{p_{p0} n_{n0}}{n_i^2} = \frac{kT}{q} \ln \frac{n_{n0}}{n_{p0}}$$

Rearranging this gives:

$$n_{n0} = n_{p0} e^{qV_{bi}/kT} \tag{4.7}$$

and

$$p_{p0} = p_{n0} e^{qV_{bi}/kT} \tag{4.8}$$

Because of the second assumption, these equations remain valid when we change the net potential. Thus:

$$n_n = n_p e^{q(V_{bi}-V_F)/kT} \tag{4.9}$$

with n_n and n_p the non-equilibrium electron densities at the boundaries of the space-charge region at n- and p-sides, respectively. Substituting 4.7 in 4.9 yields the electron density at the boundary of the depletion region on the p-side ($x = –x_p$):

$$n_p = n_{p0}\, e^{qV/kT} \tag{4.10}$$

and similarly:

$$p_n = p_{n0}\, e^{qV/kT} \tag{4.11}$$

where V can be both V_F or V_R, namely the externally applied voltage across the junction. We can also write this as:

$$\begin{aligned} n_p - n_{p0} &= n_{p0}(e^{qV/kT} - 1) \\ p_n - p_{n0} &= p_0(e^{qV/kT} - 1) \end{aligned} \tag{4.12}$$

Note that the minority carriers at the boundaries of the space-charge region increase substantially above their equilibrium under forward bias. Hence, there is an injection of minority carriers at the depletion region.
In the neutral n-region, there is no electric field, so the steady-state continuity equation 3.13 reduces to:

$$\frac{\partial p_n}{\partial t} = D_p \frac{d^2 p}{dx^2} - \frac{p_n - p_{n0}}{\tau_p} = 0 \tag{4.13}$$

Solving this equation with boundary conditions of eq. 4.12 and $p_n(x = \infty) = p_{n0}$ gives:

$$p_n - p_{n0} = p_{n0}(e^{qV/kT} - 1) e^{-(x-x_n)/L_p} \tag{4.14}$$

with $L_p = \sqrt{D_p \tau_p}$ the diffusion length of holes. This graph is shown in the lower left part of figure 4.3. At the boundary $x = x_n$:

$$J_p(x_n) = -qD_p \left.\frac{dp_n}{dx}\right|_{x=x_n} = \frac{qD_p p_{n0}}{L_p}(e^{qV/kT} - 1) \tag{4.15}$$

By applying the same reasoning for the n-region, we obtain a similar relation for J_n. Since the total current is the sum of both, we finally find:

$$J = J_p(x_n) + J_n(-x_p) = J_S(e^{qV/kT} - 1) \tag{4.16}$$

with J_S the saturation current:

$$J_S = \frac{qD_n p_{n0}}{L_p} + \frac{D_n n_{p0}}{L_n} \tag{4.17}$$

Equation 4.16 is the diode equation. Its graph is shown in figure 4.4. It is important to notice that the current increases exponentially when $V > 0$ because the potential barrier is removed. The junction will act as a conductor. On the other hand, when $V < 0$, there is only a small saturation current that is not impacted by the value of V. The junction is an open circuit (with a loss current).

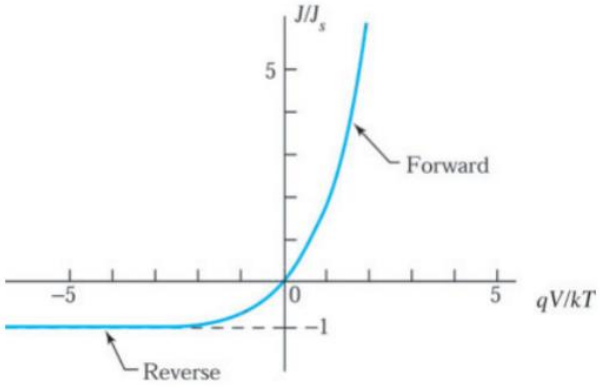


Figure 4.4: Current characteristic of equation 4.16

4.2.4 Practical Diode Characteristic

Equation 4.16 gives a current density. By multiplying with the surface A of the cross-section of the junction, we obtain the I-V curve:

$$i_D = I_S(e^{v_D/v_{th}} - 1) \tag{4.18}$$

with $v_{th} = \frac{kT}{q} \approx 26mV$ (at $T = 300K$) the thermal voltage (see figure 4.5). When $v_D \gg v_{th}$:

$$i_D \approx I_S e^{v_D/v_{th}} \tag{4.19}$$

Furthermore, as v_D increases by $\sim 60mV$, the current i_D is multiplied by a factor 10. As can be seen in figure 4.5, we can consider $v_D = 0.6V$ as a threshold voltage:

- If $v_D > 0.6V$, the diode will conduct,
- If $v_D < 0.6V$, the diode will not conduct.

Hence the diode works as a rectifier: only current from p- to n-side can pass, while the other direction is blocked. Off course, because of the saturation current and the avalanche effect (see 4.3) this is not completely true.

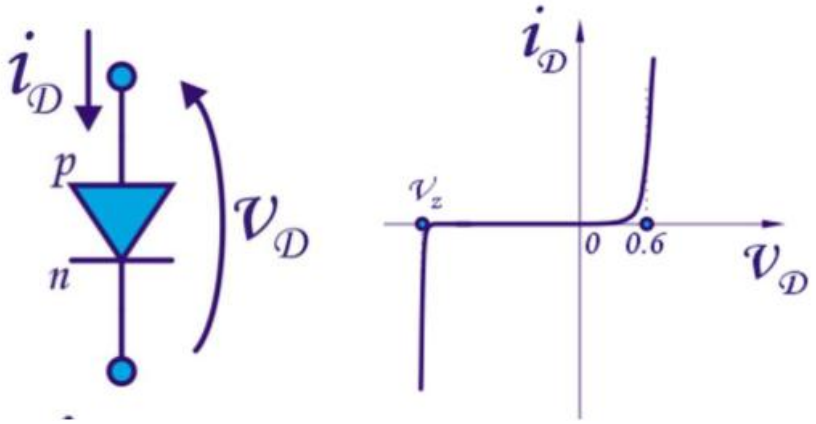


Figure 4.5: Symbol and I-V curve

4.3 Tunnel & Avalanche effect

From figure 4.5, we see that there is also a point where the diode will conduct under reverse bias. This phenomenom can have two causes: either we speak of junction breakdown due to the avalanche effect, or Zener breakdwon, due to the tunnel effect. Junction breakdown happens at high voltages and is typically unwanted. Zener breakdown can be useful and the doping is adapted such that it occurs at a couple of volts. It is for example used in voltage references (see chapter 12).

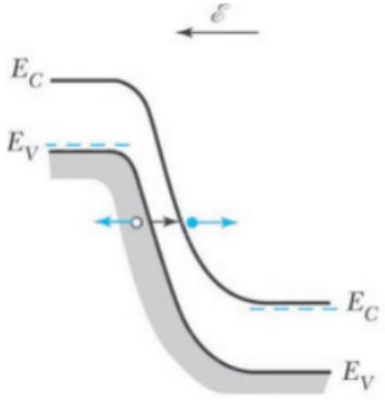


Figure 4.6

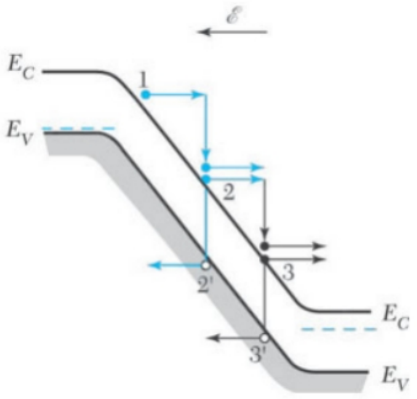


Figure 4.7

The *tunnel effect* relies on quantum tunneling: when a high electric field is applied to a p–n junction in the reverse direction, as in figure 4.6, the distance between valence and conduction band becomes locally very narrow. Under these circumstances, a valence electron can make a transition from the valence band to the conduction band. This process, in which an electron penetrates through the energy bandgap, is called tunneling. The resulting electron and hole is then swept by the electric field through the space-charge region, which creates a Zener current.

We speak of the *avalanche effect* when a thermally generated electron in the space-charge region (designated by 1 in figure 4.7) gains kinetic energy from the electric field. If the field is sufficiently high, the electron can gain enough kinetic energy that on collision with an atom,

it can break the lattice bonds, creating an electron-hole pair (2 and 2′). The newly created electron and hole both acquire kinetic energy from the field and create additional electron-hole pairs (e.g., 3 and 3′). These in turn continue the process, creating other electron-hole pairs. This process is therefore called avalanche multiplication.

4.4 Depletion Capacitance

When we reverse-bias a pn-junction, more positive charges appear on the n-side and more negative charges on the p-side. Thus, the device basically operates as a *capacitor*. In essence, we can view the conductive n and p sections as the two plates of the capacitor. We also assume the charge in the depletion region equivalently resides on each plate.
But there is more: as V_R increases, so does the width of the depletion region. That is, the capacitance of the structure decreases as the two plates move away from each other. The junction therefore displays a voltage-dependent capacitance. We can show that the junction capacitance C_j is given by:

$$C_j = \frac{C_{j0}}{\sqrt{1 - \frac{V_R}{V_{bi}}}} \tag{4.20}$$

with C_{j0} the capacitance under zero external bias ($V_R = 0$).
A pn-junction under reverse bias is thus a voltage-controllable capacitor. This kind of device has many uses, like e.g. in frequency-tunable circuits.

4.5 Photodetector

When we reverse-bias a pn-junction made of a direct-bandgap semiconductor like GaAs, we make a photodetector, as in figure 4.8. If a photon impinges on an electron in a covalent bond, it can break this bond if the energy $h\nu$ is higher than the bandgap energy. The resulting electron-hole pair is then swept across the space-charge region by the applied electric field. They recombine in the neutral zones, generating a current proportional to the number of photons that impinge on the junction.

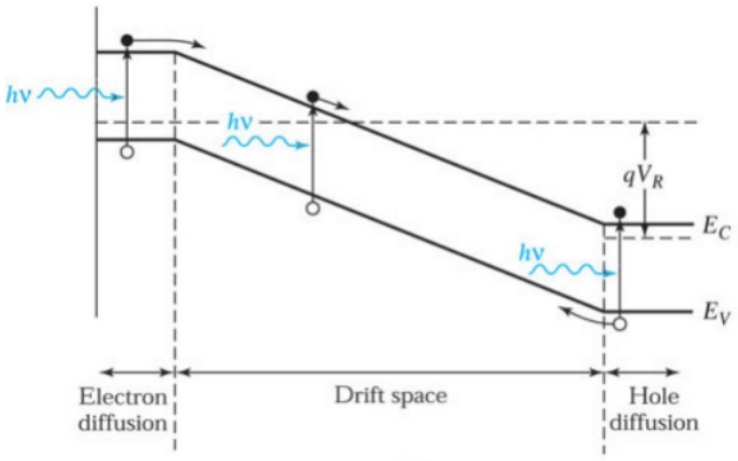


Figure 4.8: Photodetector operation

Chapter 5

Transistors

A transistor is a semiconductor device used to amplify or switch electrical signals and power. It is a component with (at least) three terminals. A voltage or current applied to one of the terminals controls the current through another pair of terminals. Because the controlled (output) power can be higher than the controlling (input) power, a transistor can amplify a signal.
We will discuss the two most prevalent transistors: the Bipolar Junction Transistor (BJT) and the Metal-Oxide-Semiconductor Field Effect transistor (MOSFET).

5.1 Bipolar Junction Transistor

A bipolar junction transistor (BJT) is formed by placing a n-type semiconductor between two p-type semiconductors (pnp) - or vice versa (npn), as shown in figure 5.1. In the case of the pnp transistor, the heavily doped p^+-region is called the *emitter* (E). The narrow central n-region is the *base* (B). Its width is small compared with the diffusion length of the minority carriers. The lightly doped p-region is the *collector* (C). Figure 5.1 also shows the circuit symbols for the two BJT transistors, together with the conventional directions for voltages and currents.
We will discuss the pnp-transistor in some detail; the reasoning for the npn-transistor is similar and left as an exercise.

5.1.1 Operation in Active Mode

First consider the pnp-transitor with the three terminals (emitter, base and collector) connected to ground as in figure 5.2 (a). Under these circumstances we can construct the charge profile like we did for the pn-junction. Notice that the space-charge regions at both junctions extend deeper in the lower doped regions (base in E-B junction, collector in B-C junction).
As for the pn-junction, the electric field that emerges in the space-charge region is such that it in equilibrium with the diffusion current due to the concentration gradients at the junctions. Figure 5.2(d) shows the band diagram in equilibrium. Since there is no current, the Fermi-level E_F is constant in the three regions.
We will polarize the transistor in the so-called active mode. This is the most commonly used mode in practice. We do this by applying a positive voltage $V_{EB} > 0$ between emitter and base, and a negative voltage $V_{CB} < 0$ between collector and base. The pn-junction between

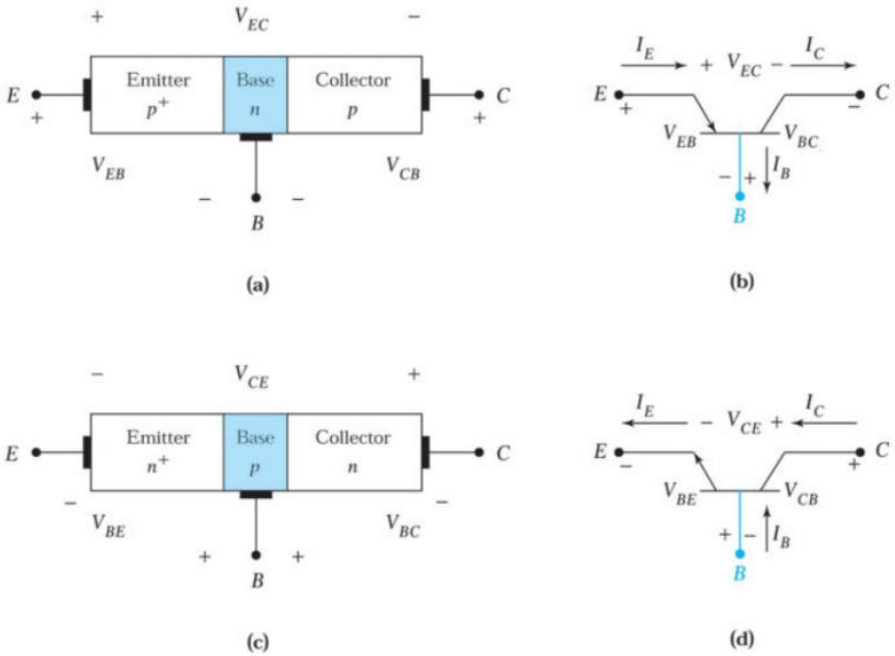


Figure 5.1: (a) Idealized one-dimensional schematic of a p-n-p bipolar transistor and (b) its circuit symbol. (c) Idealized one- dimensional schematic of an n-p-n bipolar transistor and (d) its circuit symbol.

emitter and base is thus forward-biased, while the junction between collector and base is reversed biased.
Since we apply a positive voltage to the emitter and a negative voltage to the collector (while keeping the base grounded; this is a *common-base* configuration), we lower the energy bands in the emitter and raise them in the collector (just as in the pn-junction) - see figure 5.3(d). As a consequence, the potential barrier between emitter and base is lowered so majority carriers of the emitter side (i.e. holes) can diffuse through the space-charge region to the base. If the base would be a lot larger than the diffusion length, almost all injected holes would recombine with electrons in the base and generate a base current, just like in an ordinary pn-junction. However, since the base is small, most injected holes do not recombine but diffuse into the space-charge region between base and collector. There they are swept by the electrical field to the collector. As a consequence, most holes leaving the emitter end up in the collector. This phenomenon is called the *transistor action*. It is important to realize that the collector current depends on the emitter current and thus on the height of the emitter-base barrier and V_{EB}, but not on the base-collector voltage V_{CB} (as long as the base-collector junction is reversed biased).

5.1.2 Currents in Active Mode

For each terminal, we can identify the carrier flows that contribute to the currents I_E, I_C and I_B.

1. At the emitter:

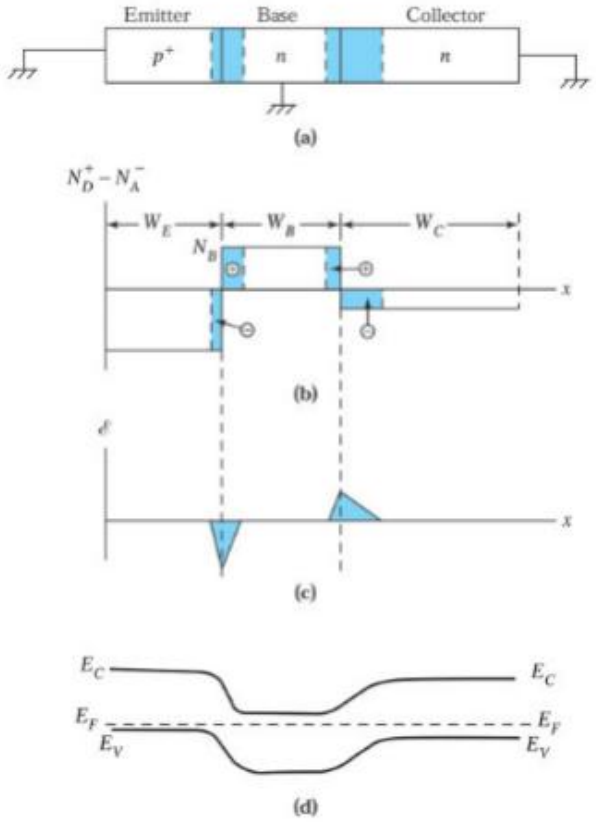


Figure 5.2: (a) pnp transistor with all leads grounded (b) Charge profile in equilibrium (c) Electric field (d) Energy band diagram.

- a hole current from emitter to the base I_{Ep}
- an electron flow from base to emitter I_{En}

Thus $I_E = I_{Ep} + I_{En}$

2. At the base:

   - a current I_{BB} due to recombination of injected holes with electrons (which have to be resupplied by the battery). This current equals the difference between I_{Ep} and I_{Cp}: $I_{BB} = I_{Ep} - I_{Cp}$
   - an electron flow from base to emitter I_{En}
   - an electron flow from collector to base: I_{Cn} (i.e. the leakage current from the reversed biased B-C junction)

   Thus $I_B = I_{En} + (I_{Ep} - I_{Cp}) - I_{Cn}$

3. At the collector:

   - a current I_{Cp} that is what remains of I_{Ep} after transition through the base.

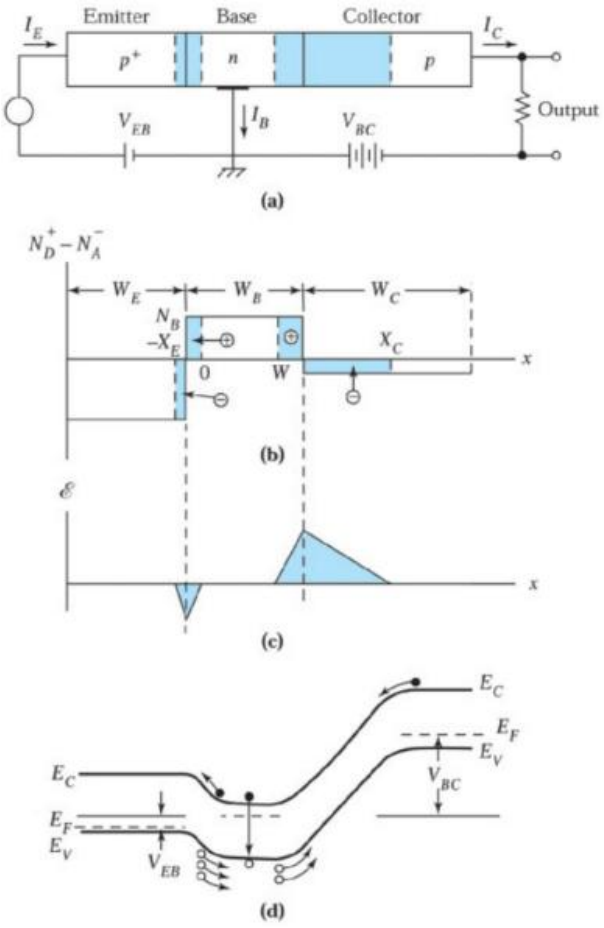


Figure 5.3: pnp transistor in active mode

- an electron flow from collector to base through the reversed biased B-C junction: I_{Cn}

Thus $I_C = I_{Cp} + I_{Cn}$

These currents are represented in figure 5.4. As the figure implies, I_{Ep} is the major current in the device.
We characterize the transistor by the *common-base current gain* α_0, which is the ratio between I_{Cp} and the emitter current:

$$\alpha_0 \equiv \frac{I_{CP}}{I_E}$$

We can rewrite this as:

$$\alpha_0 = \frac{I_{Cp}}{I_{En} + I_{Ep}} = \left(\frac{I_{Ep}}{I_{En} + I_{Ep}}\right)\left(\frac{I_{Cp}}{I_{Ep}}\right) = \alpha_T\, \gamma$$

with α_T the *base transfer factor* and γ the *emitter efficiency*. For proper operation, we require that $I_{Ep} \gg I_{En}$, thus that $\alpha_T \approx 1$. This can be accomplished by requiring that the emitter doping level be much greater than that of the base. The factor γ can be increased by

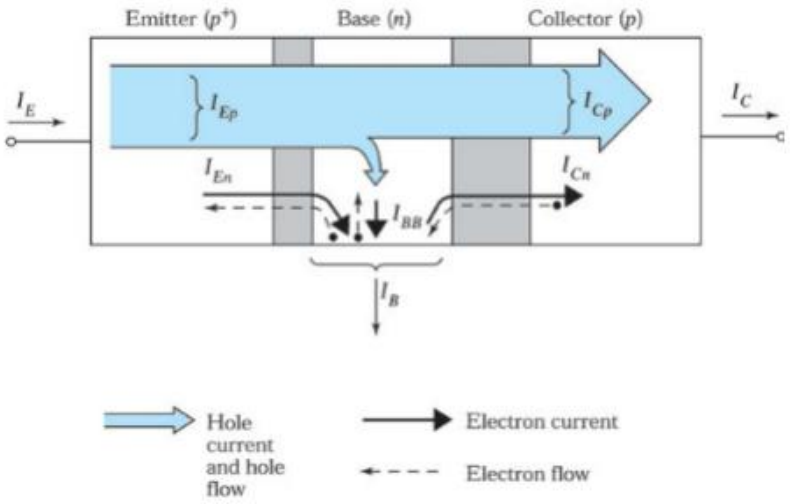


Figure 5.4: Currents in active mode

decreasing the length of the base.
We can rewrite the collector current I_C as:

$$\begin{aligned} I_C &= I_{Cp} + I_{Cn} = \alpha_T I_{Ep} = \alpha_0 I_E + I_{Cn} \\ &= \alpha_0 I_E + I_{CB0} \end{aligned} \tag{5.1}$$

where I_{CB0} is the leakage current between collector and base with the emitter-base junction open. This expresses that the collector current is a fraction ($0 < \alpha_0 < 1$) of the emitter current, plus a leakage current.

5.1.3 Carrier distribution in Active Mode

In order to compute the different currents, we first need to determine the carrier distribution in each region. We will assume the following:

1. Uniform doping in each region
2. No hole drift current in base
3. Collector saturation current is negligible
4. Only low-level injection
5. No generation-recombination in the depletion zone
6. No series resistance in the device

We will assume that the $x = 0$ corresponds to the end of the emitter-base depletion zone, and $x = W$ to the start of the base-collector depletion zone. Starting from the continuity equation for minority carriers in the base region:

$$\frac{dp_n}{dt} = -\frac{(p_n - p_{n0})}{t_p} - \frac{1}{q}\frac{dJ_p}{dx} + g \text{ with } J_p = q\mu_p pE + qD_p\frac{dp_n}{dx}$$

If we assume steady-state, no electric field and no other generation mechanisms, the equation reduces to:

$$D_p \frac{d^2 p_n}{dx^2} = \frac{(p_n - p_{n0})}{t_p}$$

This equation has as general solution

$$p_n(x) = p_{n0} + C_1\, e^{x/L_p} + C_2\, e^{-x/L_p}$$

with $L_p = \sqrt{D_p t_p}$ the so-called *diffusion length* of the holes in the n-type semiconductor. Constants C_1 and C_2 can be determined by the boundary conditions:

- $p_n(0) = p_{n0}\, e^{qV_{EB}/kT}$ because this the standard concentration at the boundary of the depletion region, as in equation 4.11 of chapter 4,
- $p_n(W) = 0$ because all holes at $x = W$ will be swept through the base-collector space-charge region by the induced electric field.

Solving for C_1 and C_2 leads to a rather complex distribution. However, if $W/L_p \ll 1$ we can simplify and obtain:

$$p_n(x) = p_{n0}\, e^{qV_{EB}/kT} \left(1 - \frac{x}{W}\right)$$

The minority carrier concentration decreases thus linearly in the base. In a similar manner, we can find an expression for the minority carriers (electrons) in emitter and collector. The results are:

$$\begin{aligned} n_E(x) &= n_{E0} + n_{E0}(e^{qV_{EB}/kT} - 1)e^{(x+x_E)/L_E} \\ n_C(x) &= n_{C0} - n_{C0}e^{(x-x_C)/L_C} \end{aligned} \tag{5.2}$$

These distributions are also represented in figure 5.5. Once the carrier concentrations are known, the currents are easily computed as they are proportional to the concentration gradient:

- $I_{Ep} = A\Big(-qD_p \frac{dp_n}{dx}|_{x=0}\Big) = \frac{qAD_p\, p_{n0}}{W} e^{qV_{EB}/kT}$
- $I_{En} = A\Big(-qD_E \frac{dn_E}{dx}|_{x=-x_E}\Big) = \frac{qAD_E\, n_{E0}}{L_E}(e^{qV_{EB}/kT} - 1)$
- ...

By combining these results, we obtain expressions for the three terminal currents:

- $I_E = a_{11}(e^{qV_{EB}/kT} - 1) + a_{12}$
- $I_C = a_{12}(e^{qV_{EB}/kT} - 1) + a_{22}$
- $I_B = I_E - I_C$

where all the a_{ij} depend on the transistor characteristics like doping concentrations and dimensions.

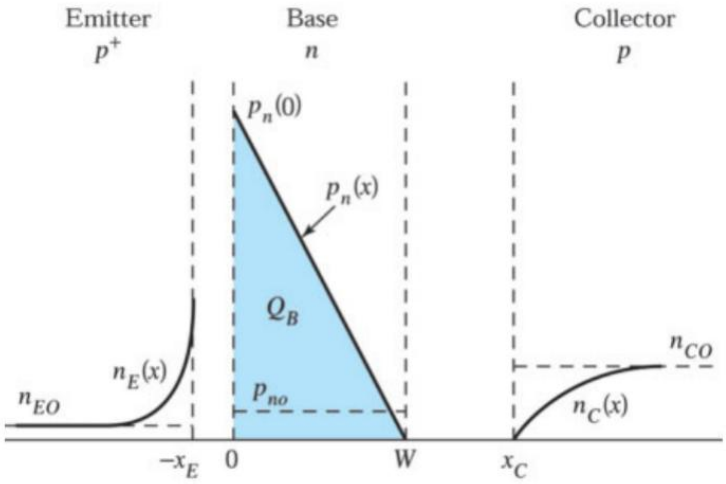


Figure 5.5: Minority carrier distribution in active mode

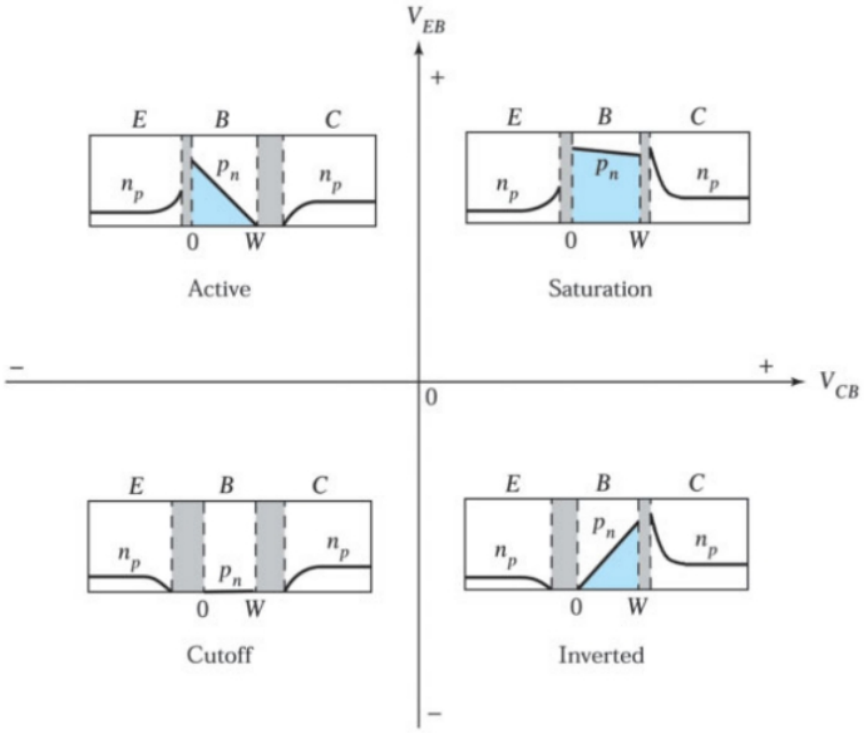


Figure 5.6: Modes of operation

5.1.4 Modes of Operation and Ebers-Moll Equations

Based on the signs of V_{EB} and V_{CB}, we can distinguish 4 different modes of operation. We already studied the case where $V_{EB} > 0$ and $V_{CB} < 0$, namely the active mode. The other modes are represented in figure 5.6, together with the minority carrier distribution in emitter, base and collector. In saturation mode, both junctions are forward biased. The minority-carrier distribution at the edge of each depletion region is not zero, as in the active mode.

Small biasing voltages lead to a large output current. The transistor acts a closed switch.
In cuttoff mode, both junctions are reversed biased. Only a small current will flow. The transistor is an open switch.
In inverted mode, emitter and collector are reversed. The behavior is however not exactly like in the active mode, because the doping levels in emitter and collector are different.
All these modes can be analyzed in a similar way as we have done for the active mode. This analysis leads to the *Ebers-Moll equations* which are used to model the BJT in simulators like SPICE.
TODO: Add Ebers-Moll equations and equivalent circuits

5.1.5 Common-Emitter configuration

Up until now, we have kept the base at a fixed voltage. However, the most common use of the bipolar transistor is with the emitter at a fixed voltage as in figure 5.7(a). This configuration has V_{EB} and I_B as inputs, and I_C and V_{EC} as outputs.
We can express I_C as function of I_B by substituting $I_E = I_B + I_C$ in equation 5.1:

$$I_C = \alpha_0(I_B + I_C) + I_{CBO}$$

Solving for I_C gives:

$$\begin{aligned} I_C &= \frac{\alpha_0}{1-\alpha_0} I_B + \frac{I_{CB0}}{1-\alpha_0} \\ &= \beta_0 I_B + I_{CE0} \end{aligned} \tag{5.3}$$

with $\beta_0 = \frac{\alpha_0}{1-\alpha_0}$ the *common-emitter current gain* and $I_{CE0} = \frac{I_{CB0}}{1-\alpha_0}$ the collector-emitter leakage current. This current is always present, even if $I_B = 0$ and thus makes the bipolar transistor a very bad switch because it will leak current when closed.
In the previous, common-base configuration, and with $\alpha_0 \approx 1$, a change in emitter current I_E produces a changes of approximately the same amount in the collector current I_C and a much smaller change in the base current (factor of $1 - \alpha_0$). To achieve current amplification, the change is initiated in the base current rather than in the emitter current. This causes the collector current to change by a factor of $\frac{\alpha_0}{1-\alpha_0} = \beta_0$.
When α_0 is close to one, β_0 becomes very large. This is a good thing because it allows us to amplify the base current I_B. However, β_0 can change a lot from transistor to transistor. This is why we will explore polarisation techniques that mitigate the variability of β_0.
The defining I-V characteristic of a common-emitter configuration is shown in 5.7(b). If V_{EC} is too low, V_{CB} is positive and the base-collector junction is not reversed biased. The transistor in saturation mode. Thus V_{EC} must be higher than some threshold to put the transistor in the active mode. This value is called $V_{EC,Sat}$ and is about 0.2 V.
When $V_{EC} > V_{EC,Sat}$, we can say based on equation 5.3 that $I_C \approx \beta_0 I_B$ and hence independent of V_{EC}. This is the ideal behavior, because we now control output current I_C with input current I_B and not with the output voltage V_{EB}. However, as can be seen from the figure, the I-V characteristics are not entirely flat and thus still depend on V_{EB}. This is due to the *Early-effect* and will be addressed in the next section.
If V_{EB} is lower than 0.6 V, the emitter-base junction is not forward biased and I_B is very small. We say that the transistor is in cut-off. The only current that still flows from emitter to collector is the leakage current I_{CE0}.

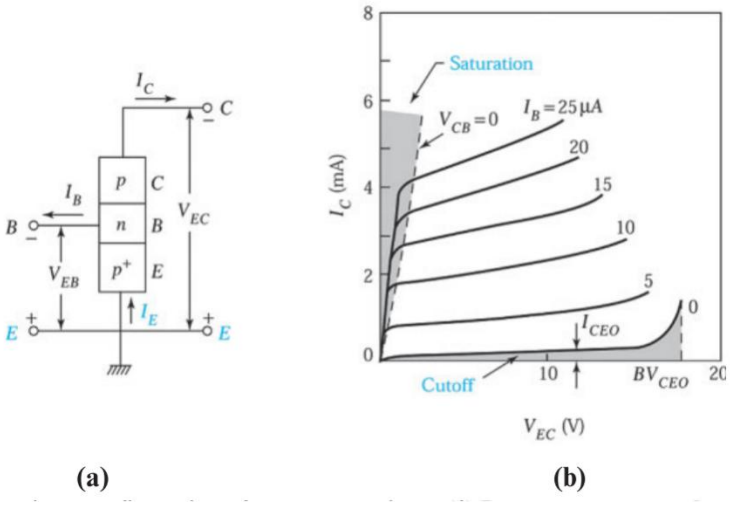


Figure 5.7: BJT in common-emitter configuration (a) and I-V characteristic (b)

5.1.6 Early Effect

In principle, I_C should be independent of the base-collector junction voltage V_{BC}. However, as V_{BC} increases, the width of the space-charge region between base and collector increases, as predicted by equation 4.6 where $V_F = -V_{BC}$. This means that the effective length of the base decreases and this in turn has two effects:

1. More holes coming from the emitter will reach the collector because there is less room for recombination. Effectively, the base transfer factor α_T increases.

2. The hole current is determined by the slope of the hole concentration in the base, as seen previously. As W decreases - while the boundary conditions remain the same - the diffusion hole current I_{Ep} increases. The increase of slope (in absolute terms) is shown in figure 5.8(a).

Both effects contribute to an increase in I_C when V_{CB} (or V_{EC} in common-emitter) increases and I_B remains constant. This phenomenon is called the *Early effect*. We can show that all those non-flat current curves in the I-V characteristic (see 5.8(b)) pass through the same point on the V-axis when extended. The corresponding voltage is the Early voltage V_A.

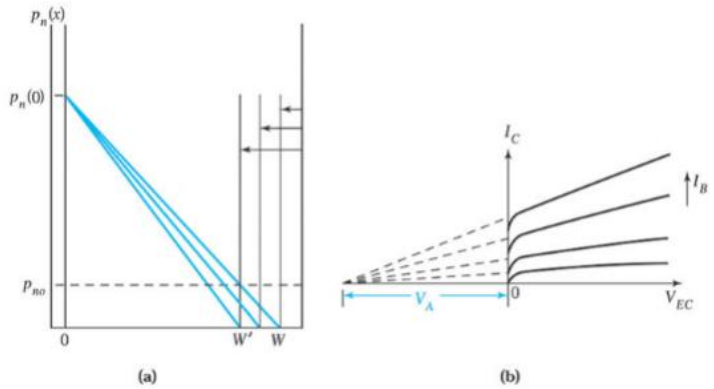


Figure 5.8: Schematic diagram of (a) the Early effect and (b) Early voltage V_A

5.2 MOSFET

The metal–oxide–semiconductor field-effect transistor or *MOSFET* is a transistor based on the field-effect. It has an insulated gate, made of a conductor like a metal or poly-silicon. This terminal is isolated from the rest of the device by a narrow layer of SiO_2 which serves as dielectric. The voltage applied at the gate determines the conductivity of the device and hence can control the current between the two other terminals named source and drain. This ability to change conductivity with the amount of applied voltage can be used for amplifying or switching electronic signals. Figure 5.9(a) shows the structure of a MOSFET, while Figure 5.9(c) shows its symbols together with the three terminals.

5.2.1 Description and Operation

Figure 5.9(b) shows the cross-section of an n-channel MOSFET. Source (S) and drain (D) are both n^+ contacts and are isolated from each other by the p-type substrate. As the voltage on the gate increases, the holes in the p-type material are expelled and the region below the gate becomes depleted of charge carriers. If the gate voltage increases even further, electrons which are present in the p-type bulk as minority carriers are attracted to the region below the gate. A narrow layer of electrons between source and drain is formed and conduction of electrons between source and drain becomes possible. If the electron density below the gate is equal to the hole concentration in the bulk, inversion has happened and we say that a channel has formed. The gate voltage at which this inversion from p-type to n-type happens, is called the threshold voltage V_T.
To have a current flow from source to drain, we don't only need a channel but also a positive voltage difference between drain and source. As we increase the voltage at the drain, the current increases, but at the same time the depth of the channel at the drain decreases because $V_{GD} < V_{GS}$. At a certain moment the voltage difference between gate and drain is no longer enough to sustain an n-type channel ($V_{GD} = V_T$). We call this *pinch-off* and say that the transistor is saturated. The current between drain and source no longer increases as the drain voltages increases but remains constant.
The dimensions of a MOSFET can become very small; a typical gate oxide thickness t_{ox} is about $15\mathring{A}$ and shrinks with every new generation. The channel length L is typically multiple

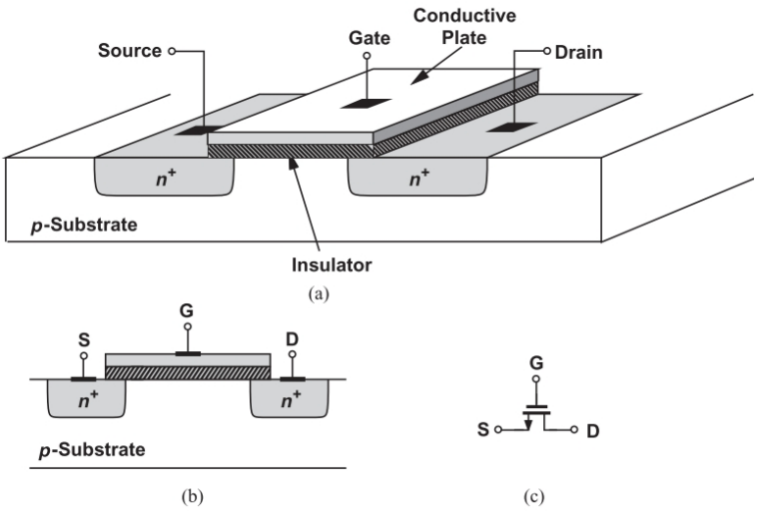


Figure 5.9: Schematic diagram of a n-channel MOSFET: (a) structure, (b) side view, (c) symbol

tens of nanometers.

5.2.2 MOS Capacitor

The region below the gate is a called a MOS capacitor. In integrated circuits, it can store charges and forms the basis building block for charge-coupled devices (CCD). A MOSFET can thus be seen as a MOS capacitor and two pn-junctions placed immediately adjacent to it. We will briefly see how charge builds up in the semiconductor as a voltage is applied to the metal gate.
Figure 5.10 represents a MOS capacitor with the gate on the left and the p-type semiconductor on the right of the oxide. If a negative voltage $V < 0$ is applied to the metal plate of the gate, as in 5.10(a), positive carriers are attracted to the $SiO_2 - Si$ interface. No current flows in the device, so the Fermi level remains constant. The carrier distribution depends on the difference $E_i - E_F$: $p_p = n_i e^{(E_i - E_F)/kT}$ so the conduction and valence bands at the interface have to bend upward to increase $E_i - E_F$ because E_F is constant. The holes "float" up to the maximum in the valence band and accumulate at the interface.
If $V > 0$, as in 5.10(b), holes are repelled from the interface and the bands bend up. Initially, all acceptor donors are exposed and a charge layer of depth W is created inside the semiconductor. The induced charge density per unit area is $Q_d = qN_AW$. We call this process *depletion*. If V increases further and the bands bend even more, the intrinsic Fermi level will fall below the true Fermi level as in 5.10(c) and electrons will swarm to the interface because the exponent in expression $n_p = n_i e^{(E_F - E_i)/kT}$ becomes positive. This process is called *inversion* and is the condition needed to form a channel in a MOSFET. The channel is a highly charged region just right of the interface (see the charge distribution), separated from the p-type semiconductor by a relatively wide depletion region. The voltage where the Fermi levels cross is the threshold voltage V_T and we denote the associated charge Q_n.

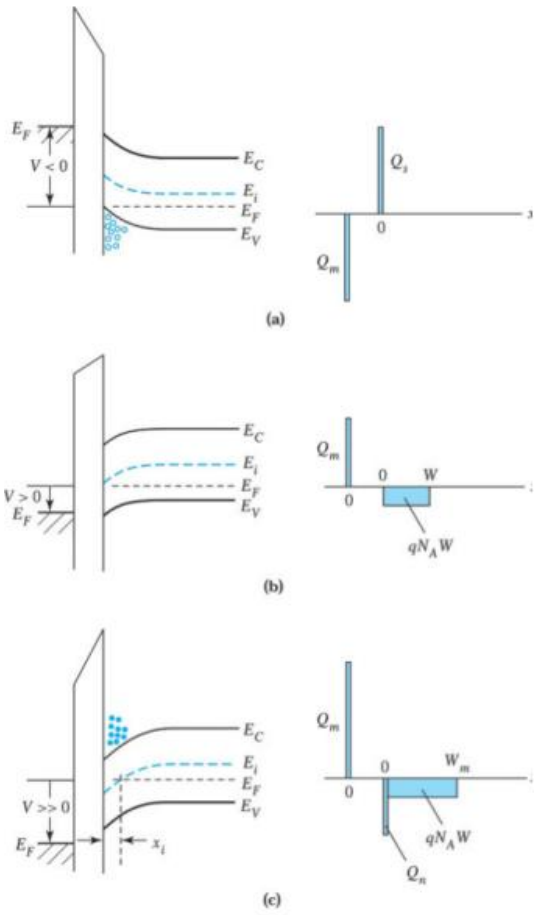


Figure 5.10: Band diagrams and charge distribution of a MOS capacitor in (a) accumulation, (b) depletion, and (c) inversion

5.2.3 I-V characteristic

If V is the applied voltage difference to the plates of a capacitor, and C is the capacitance, then the induced charge is $Q = CV$. In the case of the MOSFET of figure 5.9(b), we are only interested in the mobile charges under the gate, namely the electrons Q_n attracted to the interface in figure 5.10(c), and not the depletion charge $Q_d = qN_AW$. This means that $V = V_{GS} - V_T$ because no mobile charges exist for $V_{GS} < V_T$[1]. If we assume that the MOS capacitor has a gate capacitance C_{ox} per unit area, this relation becomes

$$Q_n = WC_{ox}(V_{GS} - V_T)$$

with W the width of the transistor. Mind that Q_n is a charge density per unit length. As drain and source voltage are not the same, the channel voltage varies along the length of the channel. If we denote the channel potential by $V(x)$, we can rewrite the equation above as:

$$Q_n(x) = WC_{ox}(V_{GS} - V(x) - V_T)$$

[1]This assumption will be revised in section 5.2.4

where $V(x)$ goes from zero to V_D if the channel is not pinched off (figure 5.11(a)).
We know that the current is the charge density times the velocity of the charges: $I_D = Q_n(x) \cdot v$. The current is a drift current as we apply an electric field $\mathcal{E}$ across the channel. Thus:

$$v = -\mu_n \mathcal{E} = \mu_n \frac{dV}{dx}$$

with μ_n the electron mobility. Substituting this in the expression for I_D, we find:

$$I_D = WC_{ox}(V_{GS} - V(x) - V_T)\mu_n \frac{dV(x)}{dx}$$

Multiplying both sides by dx and integrating from $x = 0$ to the channel length L:

$$\int_{x=0}^{x=L} I_D dx = \int_{V(x)=0}^{V(x)=V_{DS}} \mu_n C_{ox} W(V_{GS} - V(x) - V_T) dV$$

Because I_D is constant along the channel, we can solve both integrals and express I_D as

$$I_D = \frac{1}{2}\mu_n C_{ox} \frac{W}{L}[2(V_{GS} - V_T)V_{DS} - V_{DS}^2] \tag{5.4}$$

This is a parabolic function that reaches a maximum for $V_{DS} = V_{GS} - V_T$:

$$I_{D,max} = \frac{1}{2}\mu_n C_{ox} \frac{W}{L}(V_{GS} - V_T)^2$$

We already established that the condition for saturation is $V_{GD} = V_T$: from this point on, the current will no longer increase as V_{DS} increases because pinch-off has occurred at the drain. Because $V_{DG} = V_{DS} - V_{GS}$ and at pinch-off $V_{DG} = -V_T$, we can rewrite this condition as $V_{DS} = V_{GS} - V_T$, i.e. at the maximum of the curve, the transistor goes in saturation. This is the situation in figure 5.11(b) and (c). As V_{DS} increases, the pinch-off point P moves closer to the source. The voltage difference in the shrinking channel at P remains $V_{GS} - V_T$. The transistor is in saturation[2] and the drain current is - in a first approximation - equal to:

$$I_D = \frac{1}{2}\mu_n C_{ox} \frac{W}{L}(V_{GS} - V_T)^2 \tag{5.5}$$

Note that contrary to the BJT, there is no DC current through the gate.[3]
If V_{DS} is relatively small, we can approximate I_D as:

$$I_D \approx \mu_n C_{ox} \frac{W}{L}(V_{GS} - V_T)V_{DS}$$

This is the expression of a resistor with value:

$$R_{on} = \frac{1}{\mu_n C_{ox} \frac{W}{L}(V_{GS} - V_T)}$$

We say that the transistor is in the *linear* or *triode* region. Since the resistance is a function of V_{GS}, the MOSFET in this region can be seen as a programmable resistance.
Figure 5.12 gives the overall output characteristic of an n-channel MOSFET. Notice how the transition point from linear to saturation depends on V_{GS}: for the transistor te be in saturation V_{DS} must be larger than the *overdrive voltage* $V_{ov} = V_{GS} - V_T$, also called the

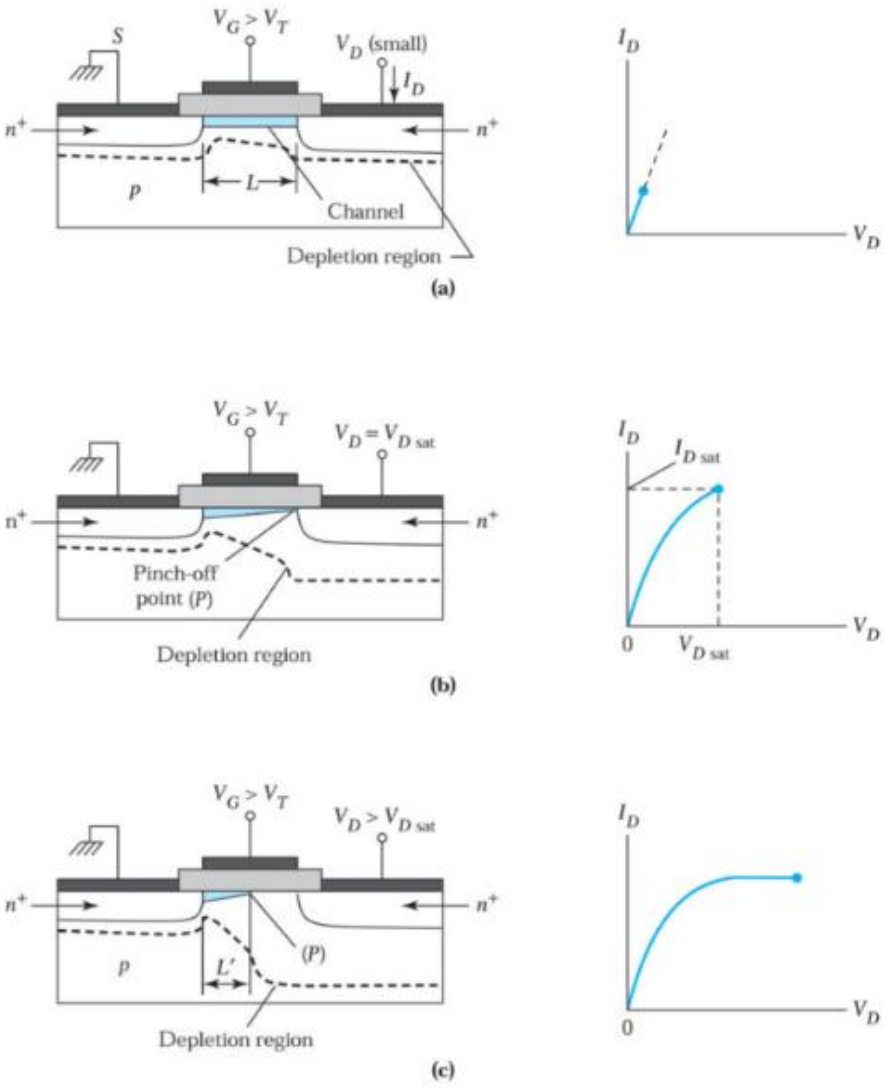


Figure 5.11: Operations of the MOSFET and output I-V characteristics. (a) Low drain voltage. (b) Onset of saturation. (c) Beyond saturation.

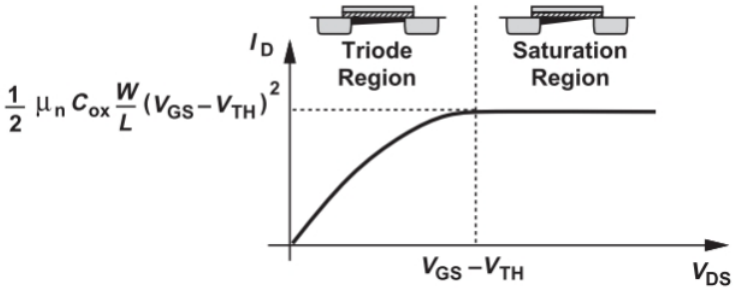


Figure 5.12: Overall MOSFET $I_D - V_{DS}$ characteristic

saturation voltage $V_{DS,Sat}$. This is not the case for the BJT, where we used a fixed cutoff $V_{CE,sat}$ with a value of 0.2 V.
We have studied the n-channel MOSFET or *NMOS*. The study of the p-channel MOSFET or *PMOS* is left as an exercise for the reader.

5.2.4 Second Order Effects

We will discuss several second-order effects that will cause the MOSFET to behave differently than the ideal behavior of figure 5.12.

Channel-length Modulation

Notice that in figure 5.12(c) the effective length of the channel decreases as the drain voltage increases (the pinch-off point moves to the left). We established equation 5.5 with the implicit assumption that L is constant. If however the effective channel length decreases, current I_D will increase with increasing V_{DS} and the I-V characteristic of figure 5.12 is not flat. This is similar to the Early effect in the BJT.
To model this, we assume that L doesn't change, but do include an explicit dependence on V_{DS} in equation 5.5:

$$I_D = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(V_{GS} - V_T)^2\,(1 + \lambda V_{DS}) \tag{5.6}$$

The factor λ is the *channel length modulation coefficient*. To decrease λ, the designer can increase the length of the transistor, because this makes the relative impact of a change in L is smaller.

Body Effect

Until now, we have supposed that both the p-type substrate and the n-type source are connected to a common ground. This is however not always the case in real circuits, where the source can be tied to voltages higher than the substrate. Even in that case, the pn-junction between source and substrate is still reversed biased and the device still works properly.
However, something does change as the source voltage increases with respect to the bulk: as the source becomes more positive with respect to the substrate, the threshold voltage V_T increases. Called "body effect," this phenomenon is formulated as

$$V_T = V_{T0} + \gamma(|2\phi_F + V_{SB} - |2\phi_F|)$$

where V_{SB} is the voltage difference between source and substrate (bulk), V_{T0} the threshold voltage when $V_{SB} = 0$ and γ and ϕ_F technology-dependent parameters.

Subthreshold Conduction

We have assumed that the MOSFET turns on abruptly when the gate voltage exceeds the threshold voltage. In practice however, the device turns on gradually and there is already a source-drain current before V_T is reached. This current depends exponentially on V_{GS}, similarly as in a BJT. Called the *subthreshold conduction*, this effect has become a critical issue in modern MOS devices.

[2] Mind that saturation for a MOSFET is not the same concept as saturation for a BJT
[3] A note on notation: we will often use K for the product $\mu_n C_{ox}$.

Part II

Analog Electronics

Chapter 6

Basic Analog Circuits

In this chapter, we will see how non-linear elements are used in electrical circuits. We discuss the concept of a load line, both static and dynamic, and study the small-signal response of a non-linear element, which essentially requires a linearization around an operating point. Finally, we provide small-signal models for both diodes and transistors at high and low frequencies.

6.1 Non-linear elements in circuits

In this section, we will see how non-linear elements like diodes and transistors are used in electrical circuits. In principle, adding a non-linear element makes the analysis of the circuit a lot harder, compared to circuits with only linear elements (resistors, capacitors, inductors, ...). In practice however, we will rely on a graphical solution method, which simplifies the analysis while still being rigorous.

6.1.1 The Diode as a Circuit Element

Let's use a diode as a lumped element in an electronic circuit, as in the figure below.
By applying KVL, we obtain the equation of the **load-line**:

$$E - v_D = R\, i_D$$

This equation has to be combined with the diode I-V characteristic:

$$i_D = \phi(v_D) = I_S(e^{v_D/v_{th}} - 1)$$

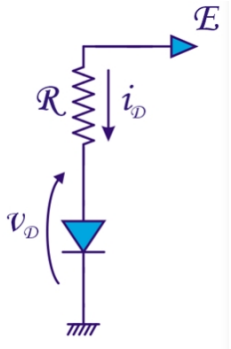


Figure 6.1

Form a formal point of view, we have two unknowns, v_D and i_d, and two equations, so in principle we can solve for both unknowns. However, there is no analytical solution to our problem, so we prefer a graphical method.
In figure 6.2 we combine $i_D = \phi(v_D)$ (the red curve) with the expression of the load line (green line).
We apply a simplification and assume that $v_D = V_{DQ} \approx 0.6$ V. The operating point[1] Q lies at the intersection of both lines. In order for the diode to conduct, we can immediately conclude

[1] The letter Q stands for *quiescent*

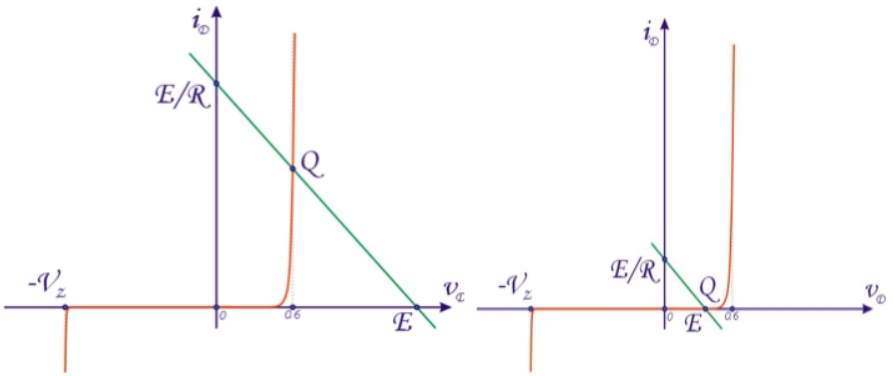


Figure 6.2: I-V with load line (a) with $E > 0.6V$ and (b) $E < 0.6$ V

by comparing both figures that it is necessary that $E > 0.6V = V_{DQ}$. The diode current is easily calculated as

$$i_D \approx I_{DQ} = \frac{E - V_{DQ}}{R}$$

We can conclude that the diode will always conduct as long as $E > 0.6$ V. Variations in E only mean that the load line will move parallel. The operating point (V_{DQ}, I_{DQ}) can only move vertically because of the nature of $\phi(v_D)$ when $v_D > 0.6$ V. Only when E becomes smaller than 0.6 V will conduction stop until we reach the Zener region. Remember that the 0.6 V is specific for silicon.

As an example, consider the circuit in figure 6.3. It is not obvious when the diode will conduct, so let's replace the current source and the resistor by the Thevenin equivalent, namely a resistor $R_{th} = R$ and a voltage source $V_{th} = RI_0$ in **series** with this resistor. This circuit is identical to the one in figure 6.1 with a load line through the points $(R\ I_0, 0)$ and $(0, I_0)$. We can conclude that the transistor will conduct when $V_{th} = R\ I_0 > 0.6$ V. Furthermore, when $R \to \infty$ and the resistor become an open circuit, the load line will become a horizontal line through I_0.

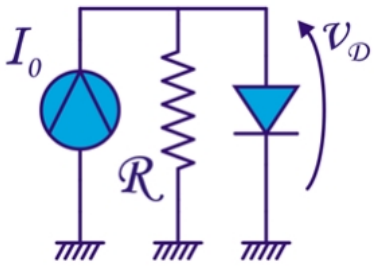


Figure 6.3: With current source

6.1.2 The BJT as a Circuit Element

A typical bipolar transistor circuit is shown in figure 6.4(b). To analyze this circuit, we cut at the entrance of the base and replace the left loop of resistors R_1 and R_2 and the supply voltage E by the Thevenin equivalent circuit in figure 6.4(a).

The Thevenin voltage E_B and impedance R_B are easily computed by seeing that (a) the voltage at the base is the result of applying a voltage divider to the supply voltage E, and

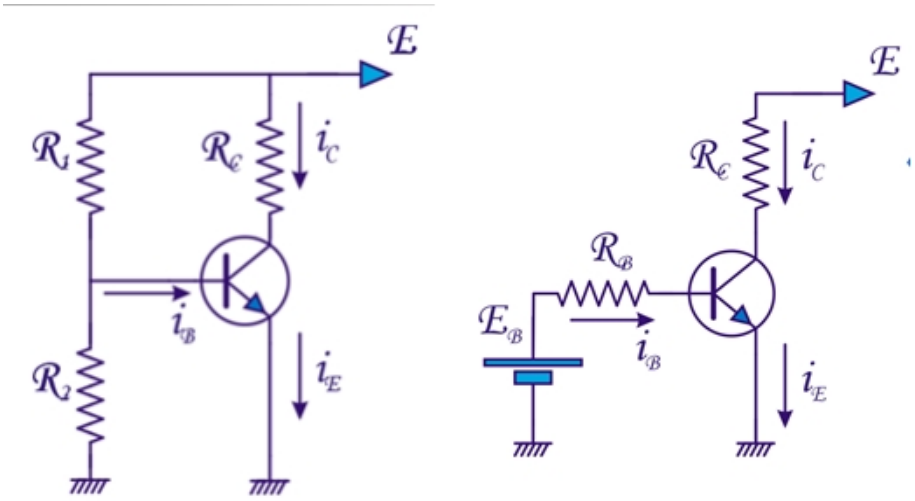


Figure 6.4: (a) Transistor circuit and (b) simplified with Thevenin

(b) that when we ground E (as required by the Thevenin rules) R_1 and R_2 are in parallel:

$$\begin{aligned} E_B &= \frac{R_2}{R_1 + R_2} E \\ R_B &= R_1 || R_2 = \frac{R_1 R_2}{R_1 + R_2} \end{aligned} \tag{6.1}$$

In this left loop, we can write:

$$E_B - v_{BE} = R_B \, i_B \tag{6.2}$$

Consider the right loop, which consists of resistor R_C and the voltage v_{CE}. In this loop, we can write:

$$E - v_{CE} = R_C \, i_C \tag{6.3}$$

Let's assume that all the currents are constant and we have biased the transistor in its operating point. We indicate this by adding a Q to the currents and voltages, just as for the diode. With this convention, we can rewrite these equations to obtain:

$$\begin{aligned} I_{BQ} &= \frac{E_B - V_{BEQ}}{R_B} \\ V_{CEQ} &= E - R_C \, I_{CQ} \end{aligned} \tag{6.4}$$

with $V_{BEQ} \approx 0.6$ V because we want to bias the base-emitter junction in the forward (conducting) region. From the diode analysis, we know this will be the case when $E_B > 0.6$ V. We also know the relation between I_{BQ} and I_{CQ} (neglecting any leak currents):

$$I_{CQ} = \beta \, I_{BQ} \tag{6.5}$$

where β is given by the manufacturer and is typically very high, but it can vary a lot from one transistor to the next. With this equation, we have enough information to compute all currents and voltages:

1. We know R_B and E_B, and we assume $V_{BEQ} = 0.6$ V, so we can compute I_{BQ}.
2. We know β, so we can compute I_{CQ}.
3. We can now compute V_{CE}.

All these calculations can all be represented graphically; see the (idealized) BJT characteristic in figure 6.5. The figure on the left represents the left loop, with the load line $E_B - v_{BE} =$

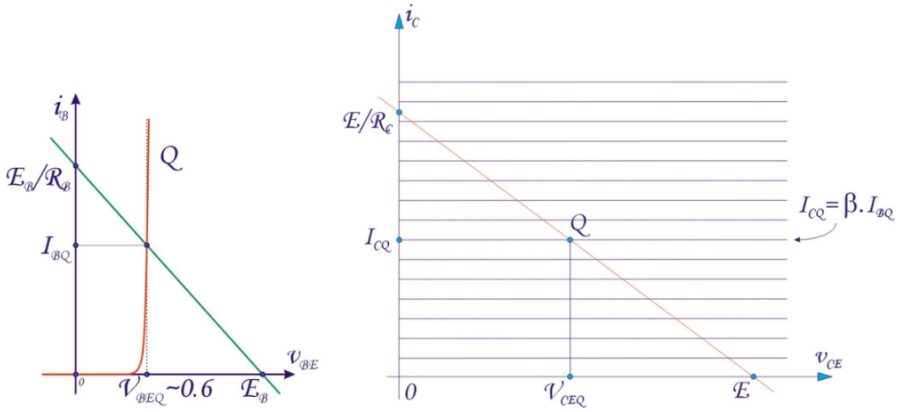


Figure 6.5: (a) Transistor circuit and (b) simplified with Thevenin

$R_B\ i_B$ in green and the v_{BE} junction characteristic in red. The intersection between both functions is the operating point $Q = (V_{BEQ}, I_{BQ})$.
The right loop is shown in the graph on the right; the horizontal line on which the transistor operates is given by $I_C = \beta\ I_B$ and the load line is given by $V_{CE} = E - R_C\ I_C$. Once again, the operating point is there where both lines intersect.
We can reason about the operation of the circuit by thinking about these figures. For example, if R_B would decrease, the slope of the $I_B - V_{BE}$ load-line would increase, so that the operating point Q will move up and I_{BQ} will increase. This increase will cause in increase of I_{CQ} in the graph on the right, and the operating point will move along the load line to a higher I_C and a lower V_{CE}.
We can conclude that:

- The left loop determines I_{BQ}.
- By assumption, we are in the normal working domain and $I_{CQ} = \beta I_{BQ}$.
- The right loop gives V_{CEQ}.
- Given $Q = (V_{CEQ},\ I_{CQ})$, we verify we are in the normal working domain.

Let's discuss a couple of edge cases:

- If $E_B < 0.6$ V then $V_{BEQ} < 0.6$ V. Then $I_{BQ} \approx 0$ and $I_{CQ} \approx 0$ (neglecting leakage current). See figure 6.6(a). If that's the case, the operating point is $Q = (V_{CEQ} = E, 0)$ and the transistor is blocked (*cut-off mode*).
- As V_{BEQ} and I_{BQ} keep increasing, I_{CQ} will become larger and eventually V_{CEQ} will become too small to keep the transistor in active mode. Then $I_{CQ} \neq \beta\ I_{BQ}$ and $V_{CEQ} \approx V_{CE,Sat}$. So, in that case, $I_{CQ} = \frac{E - V_{CE,Sat}}{R_C}$. The transistor is *saturated*.

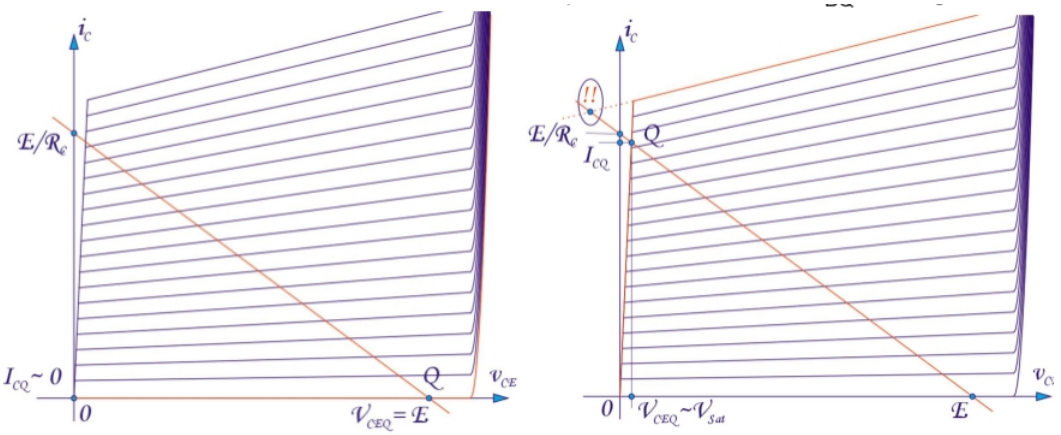


Figure 6.6: (a) $V_{BEQ} < 0.6\text{V}$ and (b) $V_{CEQ} \approx V_{CE,Sat}$

6.1.3 The MOSFET as a Circuit Element

Just as for the BJT, we consider a biasing circuit for the n-channel MOSFET transistor as in figure 6.7(a) and we simplify the circuit with a Thevenin equivalent circuit - like we did before - as in figure 6.7(b). with

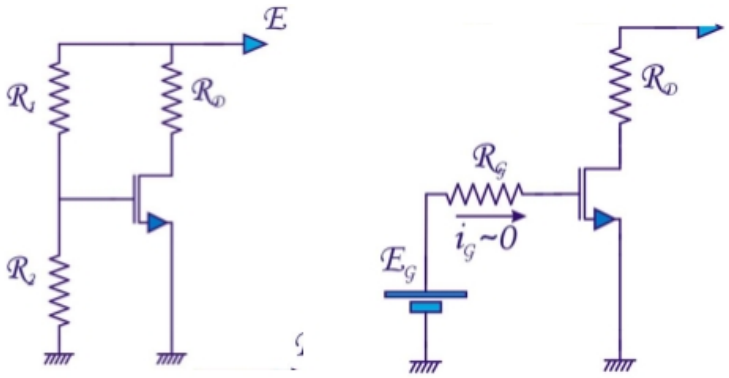


Figure 6.7: (a) MOS Transistor circuit and (b) simplified with Thevenin

$$\begin{aligned} E_B - G &= \frac{R_2}{R_1 + R_2} E \\ R_G &= R_1 || R_2 = \frac{R_1 R_2}{R_1 + R_2} \end{aligned} \tag{6.6}$$

Note that the gate current $I_G = 0$ as the gate is a capacitor where the DC current is zero. And just as for the BJT, we can express an equation for the left and right loop:

$$\begin{aligned} V_{GSQ} &= E_G \text{ because } I_G = 0 \\ I_{DSQ} &= \frac{\mu_n C_{ox}}{2} \frac{W}{L} (V_{GSQ} - V_T)^2 \text{ if } V_{GSQ} > V_T \\ V_{DSQ} &= E - R_D \, I_{DSQ} \text{ if } V_{DSQ} > V_{GSQ} - V_T \end{aligned} \tag{6.7}$$

The last condition is required for the transistor to be saturated, which is similar to the active mode for a BJT.
We can also represent these equations graphically. Figure 6.8(a) represents the quadratic relation between I_{DS} and V_{GS} to determine V_{GSQ}. As long as $V_{DS} > V_{GSQ} - V_T$, figure 6.8(b) gives the value of V_{DSQ}.

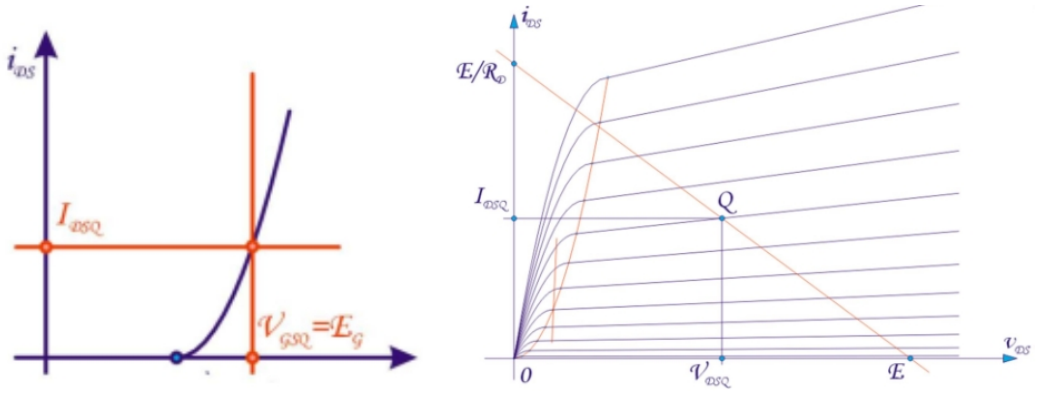


Figure 6.8: (a) $i_{DS} = f(v_{DS})$ and (b) $I_{DS} = f(V_{DS})$

6.1.4 Additional remarks

For both type of transistors, there are three working domains, as in figure 6.9:

- MOSFET: blocked, saturated, linear
- BJT: blocked, normal (or active), saturated

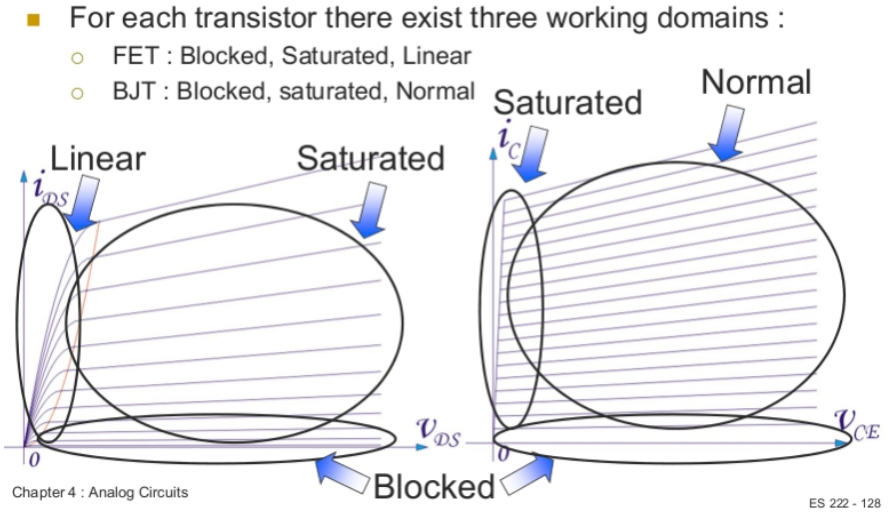


Figure 6.9: (a) $i_{DS} = f(v_{DS})$ and (b) $I_{DS} = f(V_{DS})$

6.1.5 A more general circuit

To improve the linearity (see the chapter on feedback) and biasing of the circuit, an emitter resistance R_E is often added, as in figure 6.10. Just as before, the left side is simplified with the Thevenin equivalent (see 6.11).

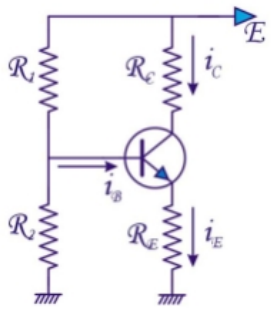


Figure 6.10: General circuit

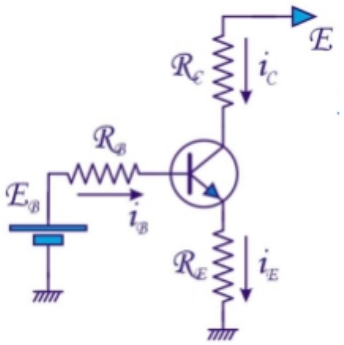


Figure 6.11: Thevenin simplification

Once again, we can write the KVL in left and right loop:

$$\begin{aligned} E_B - V_{BEQ} &= R_B I_{BQ} + R_E I_{EQ} = R_B I_{BQ} + R_E(\beta+1) I_{BQ} \\ E_B - V_{BEQ} &= R_B \frac{I_{CQ}}{\beta} + R_E(\beta+1)\frac{I_{CQ}}{\beta} \approx \frac{R_B}{\beta} I_{CQ} + R_E I_{EQ} \end{aligned} \tag{6.8}$$

where we have used $I_{CQ} = \beta I_{BQ}$ because we suppose we're working in the normal operating region. These equations lead to expressions for current I_{CQ} and voltage V_{CEQ}:

$$\begin{aligned} I_{CQ} &= \frac{E_B - V_{BEQ}}{\frac{R_B}{\beta} + R_E} \\ V_{CEQ} &= E - (R_C + R_E) I_{CQ} \end{aligned} \tag{6.9}$$

They can be plotted on the different I-V characteristics as in figure 6.12. Notice that the figure on the left gives i_C as function of v_{BE}. Since the relation between i_B and v_{BE} is the exponential diode characteristic, the same goes goes for the relation between $i_C = \beta i_B$ and v_{BE}.
We can apply the same reasoning to the 4-resistor MOSFET circuit form figure 6.13, for which we also can simplify the left loop with the Thévenin equivalent of figure 6.14. The operating point Q is determined by:

The equation of the left loop: $E_G - V_{GSQ} = R_S I_{DSQ}$
The transistor characteristic: $I_{DSQ} = \frac{K}{2}\frac{W}{L}(V_{GSQ} - V_T)^2$
The equation of the right loop: $V_{DSQ} = E - (R_D + R_S) I_{DSQ}$

To find V_{GSQ} and I_{DSQ}, use the first two equations; only one root is valid for V_{GSQ}. V_{DSQ} follows immediately from the third equation.

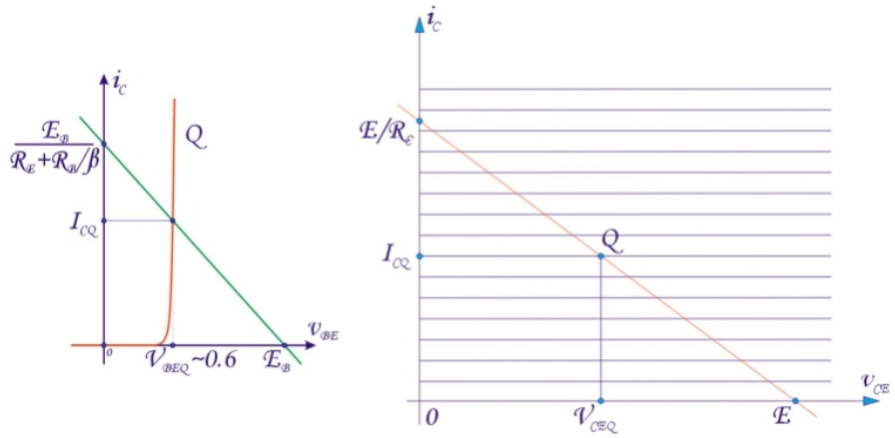


Figure 6.12: (a) Load line: $I_{CQ} = \frac{E_B - V_{BEQ}}{\frac{R_B}{\beta} + R_E}$ (b) $V_{CEQ} = E - (R_C + R_E)I_{CQ}$

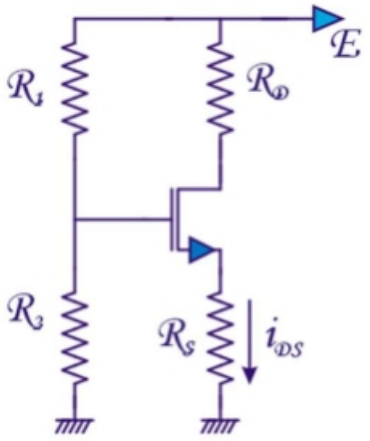


Figure 6.13

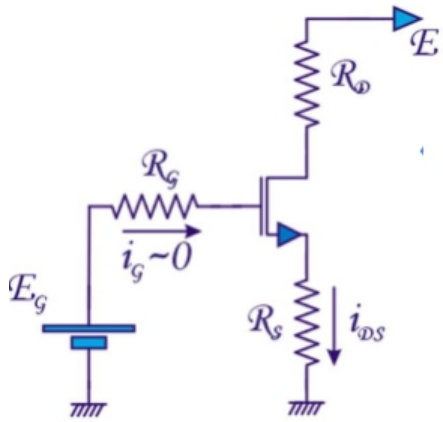


Figure 6.14

6.2 Small-Signal Response

In this section, we will introduce the concept of a small-signal response, namely how do voltage and currents in a circuit change when we apply only a small change to the input values. The general idea is that we design the circuit such that it operates at an operating point Q, and we linearize the circuit around this operating point to study only small deviations. We will use the diode as an example. in section 6.5 we apply the same reasoning to transistors.
First, we introduce some notation to distinguish large-signal from small signal quantities.

- x_A: measure of a specific variable,
- X_A: the average value of this specific variable,
- x_a: variation of the specific variable around average X_A.

Refer to figure 6.15 for a visual representation of these variables. Note that only x_A and x_a vary with time, and that the average value of x_a is zero: $\mathbb{E}[x_a] = 0$.

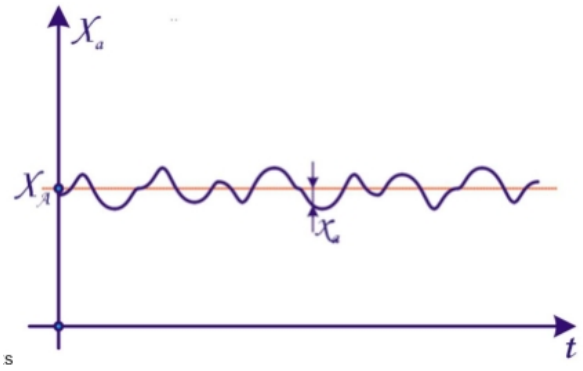


Figure 6.15: Signal quantities

Let's apply this to the simple diode circuit. In figure 6.16, the supply has two components: a fixed voltage E, and a varying voltage e with average value $\mathbb{E}[e(t)] = 0$. The quantities we're looking for, v_D and i_D, can be split in two components: an average value and a variation around this average.

$$\begin{aligned} v_D &= V_D + v_d \\ i_D &= I_D + i_d \end{aligned} \tag{6.10}$$

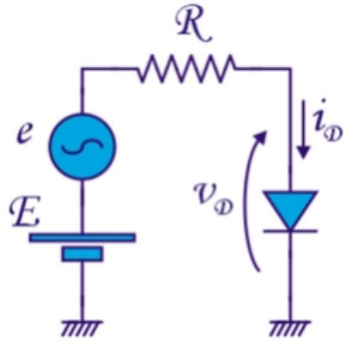


Figure 6.16

Assume that $e = 0$, i.e. we study the system with no variations. If $E > 0.6V$, we can write - as we've done before - that:

$$\begin{aligned} V_{DQ} &= 0.6V \\ I_{DQ} &= \frac{E - V_{DQ}}{R} \end{aligned} \tag{6.11}$$

In this way, we determine the operating point as the intersection between the load line and the diode characteristic, as in figure 6.17.
Now assume $e \neq 0$. This changes the equation for the load line:

$$E + e - v_D = R\, i_D \tag{6.12}$$

This equation can be rewritten as:

$$(E + e) - (V_{DQ} + v_d) = R(I_{DQ} + i_d)$$

or, since $E - V_{DQ} = R\, I_{DQ}$:

$$e - v_d = R\, i_d \tag{6.13}$$

This is the equation of the small-signal load line, where the center of the coordinate system is translated to the operating point $Q = (V_{DQ}, I_{DQ})$. Figure 6.18 shows what happens: small variations of e move the load line parallel to the original load line $E - V_{BEQ} = R\, I_{DQ}$. As this moving load line intersects with the diode characteristic, small voltage variations v_d and

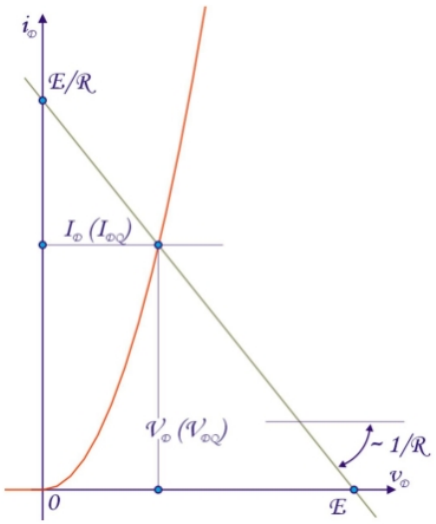


Figure 6.17: Diode equation and load line

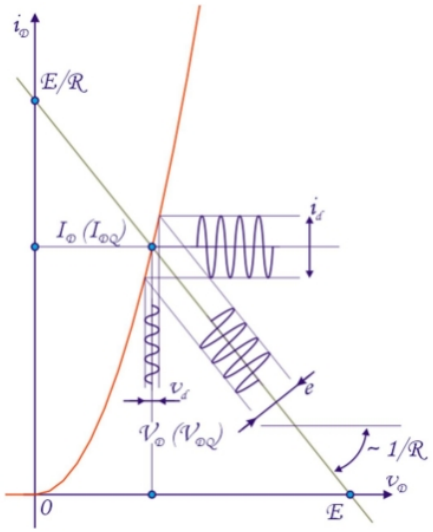


Figure 6.18: Variations around Q

current variations i_d appear across the diode. We can determine the relation between v_d and i_d:

$$\begin{aligned} i_D &= \phi(v_D) \approx I_S e^{v_D/v_{th}} \\ di_D &= \frac{I_S}{v_{th}} e^{v_D/v_{th}} dv_D \\ \Rightarrow i_d &= \frac{i_D}{v_{th}} v_d \end{aligned} \tag{6.14}$$

and finally

$$v_d = \rho_d \, i_d \text{ with } \rho_d = \frac{v_{th}}{I_{DQ}} \tag{6.15}$$

We have effectively linearized the diode characteristic around the operating point. Locally, for small variations, the diode operates as a resistor with resistance ρ_d.
By doing this, we have transformed the original problem into two subproblems:

1. Determine the DC solution by solving (graphically) the equations on the left of figure 6.19. This solution determines Q and the so-called small signal parameters, like ρ_d.

2. Solve a linear circuit where the nonlinear element has been replaced by the small-signal equivalent - a resistor in the case of our diode. See the right part of figure 6.19.

However, to be correct, the small-signal equivalent model of a diode is not just a resistor. A pn-junction creates a space-charge region on its interface. As the voltage across the junction changes, charges (both n and p) have to be transported to and from the junction to increase or decrease the SCR. This means that a diode - or any pn-junction - is also capacitive. This phenomenon of depletion capacitance was already explained in section 4.4.
To model this behavior, we replace a diode in as small-signal equivalent circuit by:

1. A dynamic resistance ρ_d, in parallel with

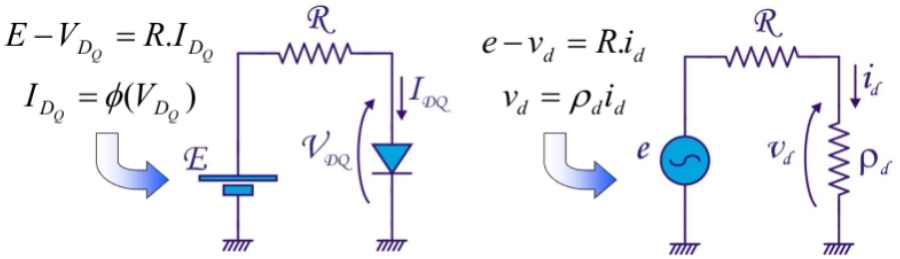


Figure 6.19: Signal quantities

2. A junction capacitance C_j, as in figure 6.20 (with $\alpha = 1/2$). This capacitance can be neglected for small frequencies.

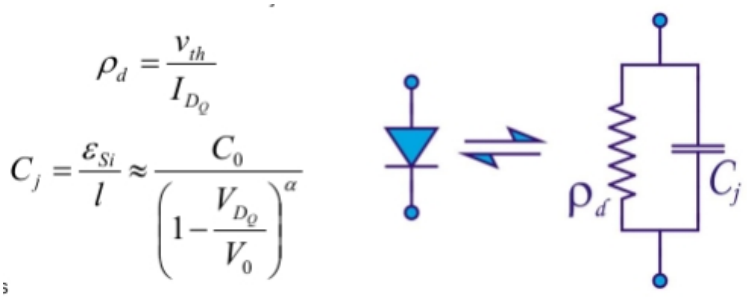


Figure 6.20: Small-signal model of a diode

Finally, to establish the small-signal equivalent circuit:

- We replace all nonlinear devices by their small-signal model (like the one in 6.20 for the diode).
- We replace all independent voltage source by a short-circuit (i.e. $E = 0$) because we assume they don't vary with time.
- For the same reason, we replace all independent current source by an open circuit (i.e. $I = 0$).

6.3 Static and Dynamic Load lines

In section 6.1, we saw the concept of a load line. However, there is more to it than we've seen up to now. The reason is the fundamental difference between the operating point Q and the small-signal response. The former is fundamentally a DC concept, because there are no time-varying quantities involved. The small-signal response at the other hand deals with AC signals: signals that vary in time and thus have a non-zero frequency. Let's thus study a circuit that contains frequency-dependent components like capacitors or inductors.
Consider the circuit in figure 6.21, where a load charge R_L is connected to the original diode circuit through a capacitor C. To compute the operating point Q, we assume $e = 0$. Since the are no variations, the capacitor C is an open circuit and the circuit is reduced to the one in figure 6.22. This is the same circuit as before, so we conclude that $V_{DQ} = 0.6$ V and the

load line is $E - V_{DQ} = R\ I_{DQ}$. The operating point allows us to compute the small-signal resistance ρ_d of the diode.

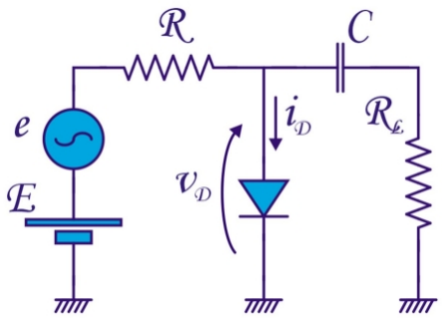


Figure 6.21: Diode circuit with capacitor

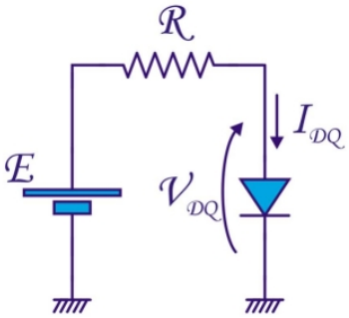


Figure 6.22: Circuit to determine Q

For the small-signal response, we can't neglect C. We do replace the diode by a resistance ρ_d (and neglect - for simplicity - the capacitance C_j) and obtain the circuit in figure 6.23 with $E = 0$. This circuit can be simplified with Thevenin's theorem, which gives the circuit in figure 6.24 where we have made a cut just above ρ_d. Z_{th} and e_{th} can be computed as follows:

- For e_{th}, we assume an open circuit, so no current through ρ_d. As such, we have a voltage divider consisting of resistors R and R_L and capacitor C. Let Z_C be the series combination of R_L and C, namely:

$$Z_C = R_L + \frac{1}{j\omega C} = \frac{1 + j\omega R_L C}{j\omega C} \tag{6.16}$$

  Then, we apply the expression for a voltage divider:

$$e_{th} = \frac{Z_C}{R + Z_C} e = \frac{\frac{1+j\omega R_L C}{j\omega C}}{R + \frac{1+j\omega R_L C}{j\omega C}} e = \frac{1 + j\omega R_L C}{1 + j\omega (R + R_L) C} e \tag{6.17}$$

- For Z_{th}, we replace e by a short-circuit. Z_{th} ts then the parallel combination of R with Z_C:

$$Z_{th} = \frac{Z_C R}{Z_C + R} = R\ \frac{1 + j\omega R_L C}{1 + j\omega (R + R_L) C} \tag{6.18}$$

Obviously, the load line is different: at DC, it is $E - V_{DQ} = R\ I_{DQ}$, but at AC it is $e_{th} - v_d = i_d Z_{th}$. For high frequencies, we can simplify $Z_{th}|_{\omega \to \infty} = R \frac{R_L}{R + R_L} = R || R_L$ and the small-signal load line becomes $e_{th} - v_d = i_d (R || R_L)$ which has a different slope than the DC load line (keep in mind that the AC load line is centered at the operating point Q).

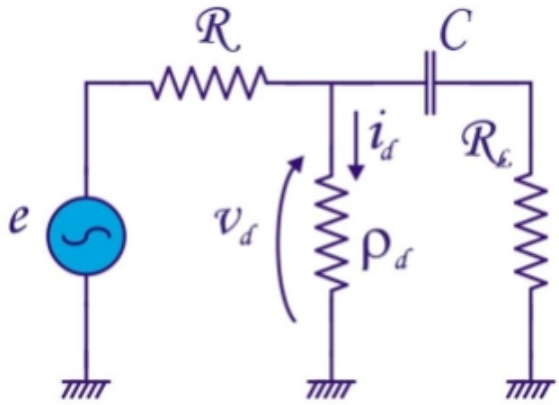


Figure 6.23: Diode circuit with capacitor

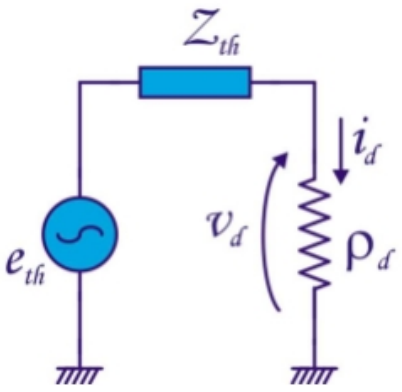


Figure 6.24: Circuit to determine Q

Both load lines are show in figure 6.25, with in green the static load line (slope $= \frac{-1}{R_{stat}}$ where $R_{stat} = R$) and in blue the dynamic load line (slope $= \frac{-1}{R_{dyn}}$ where $R_{dyn} = R||R_L$).

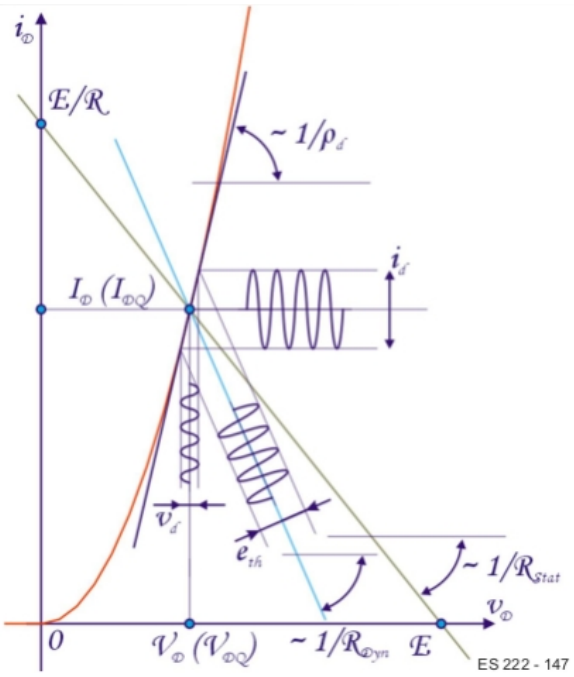


Figure 6.25: Static (green) and dynamic (blue) load lines

We conclude that there are two load lines:

1. The *static* load line, determined at zero frequency (DC) and used to place the operating point.

2. The *dynamic* load line, at the frequency of interest, which typically is high enough so that we can simplify the impedance. It determines how the operating point will move (the small-signal response).

The later remark implies that there is a critical frequency from which the capacitor C can

be neglected. From equation 6.16, we see that this impedance has a pole in $\omega = 0$ and a zero in $\omega = 1/(R_L C)$. For pulsations higher than $\frac{1}{R_L C}$, the impedance becomes frequency-independent and is equal[2] to R_L. Thus the circuit reduces to the one in figure 6.26. It is this circuit that determines the dynamic load line.

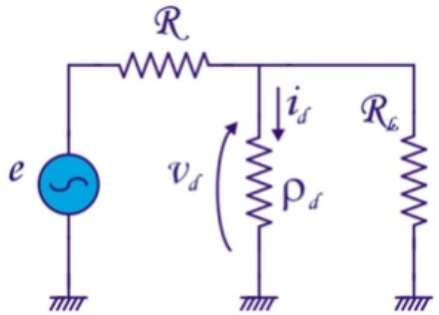


Figure 6.26: Circuit to determine the dynamic load line

6.3.1 Transistors and Dynamic Load Lines

Consider the circuit in figure 6.27. This is the same circuit as we saw in figure **??**, but with a capacitor C_E in parallel with the emitter resistance R_E. After simplifying the left part with

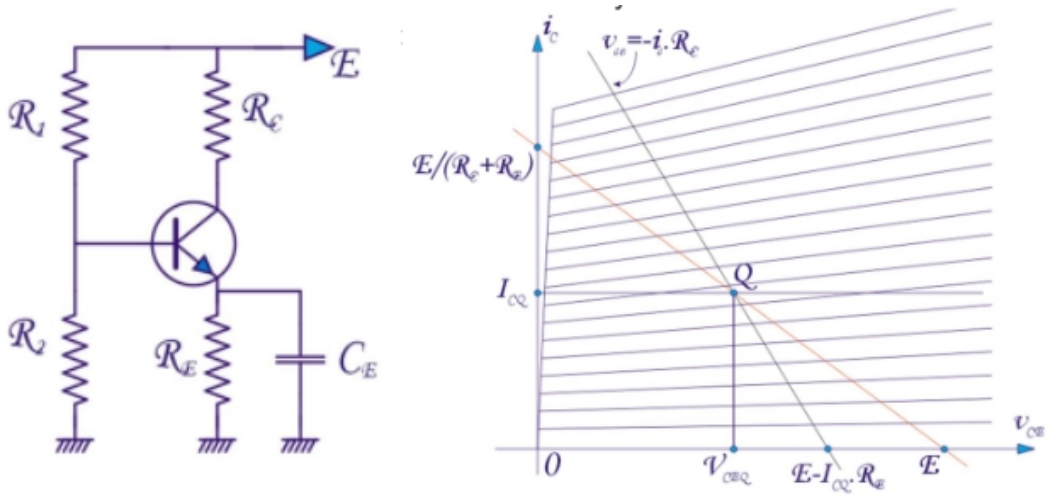


Figure 6.27: BJT circuit with emitter bypass capacitor C_E

the Thevenin theorem, we can establish the equation in the left loop, which hasn't changed from section 6.1.5:

$$E_B - V_{BEQ} = \left(\frac{R_B}{\beta} + R_E\right) I_{CQ} \tag{6.19}$$

The equation in the right loop does change, with $Z_E = R_E || C_E = \frac{R_E}{1+j\omega C_E R_E}$, and becomes

$$v_{CE} = E - (R_C + Z_E) i_C \tag{6.20}$$

[2]Convince yourself by sketching the Bode graph

which can be split into two equations:

1. The DC operating point:

$$V_{CEQ} = E - (R_C + R_E)\, I_{CQ}$$

,

2. A small signal equation, valid for $\omega \gg \omega_0$:

$$v_{ce} = -R_C\, i_c$$

These static and dynamic load lines are represented in figure 6.27 (red and green lines, respectively). The intersection of the dynamic load line with the horizontal axis is given by $E - R_E\ I_{CQ}$ because the voltage at the emitter is fixed (if ω is high enough) and equals $R_E\ I_{CQ}$. Hence, on the dynamic load line, the maximum swing of v_{ce} is between 0 and $E - R_E\ I_{CQ}$.

6.4 Biasing

In the previous section, we developed a way to determine the operating point of a transistor if the resistors are given:

- Determine the current via the left loop: $E_B - V_{BEQ} = \left(\frac{R_B}{\beta} + R_E\right) I_{CQ}$ or $E_G - V_{GSQ} = R_S I_{DSQ}$.
- Determine V_{CEQ} or V_{DSQ} based on the right loop, and check wether we are in the normal (saturation) region of the transistor (i.e. $V_{CEQ} > 0.6$ V or $V_{DSQ} > V_{GSQ} - V_T$).

However, many values are not exactly known:

- V_{BEQ} varies over a specific production lot,
- β depends on i_C, and varies (from -50% to +200%) over a specific lot,
- V_T is only specified with a certain precision,
- $K = \mu_n C_{ox}$ (or $\mu_p C_{ox}$) is only specified with a certain precision,
- All these parameters vary with temperature.

This section will describe a method to choose the biasing resistors (R_1, R_2, R_C or R_D and R_E or R_S) such that variation in the parameters above has minimal impact on the quiescent currents and voltages for which the circuit is designed.

6.4.1 BJT Biasing

The goal of BJT biasing is to choose R_1, R_2 and R_E to reduce the impact of variations on V_{BEQ} and β. Furthermore, R_C will be chosen such that the operating point Q lies in the middle of the normal operating region.
Consider the general four-resistor BJT circuit in **??**(b). We reproduce the Thevenin simplification here for convenience.
As a reminder, the Thevenin voltage E_B and impedance R_B are derived from the biasing resistors R_1 and R_2 and the supply voltage E: $E_B = \frac{R_2}{R_1+R_2}E$ and $R_B = R_1||R_2$.
From this circuit, we see that:

$$E_B - V_{BEQ} = \left(\frac{R_B}{\beta} + R_E\right) I_{CQ}$$

which means that:

$$I_{CQ} = \frac{E_B - V_{BEQ}}{\frac{R_B}{\beta} + R_E}$$

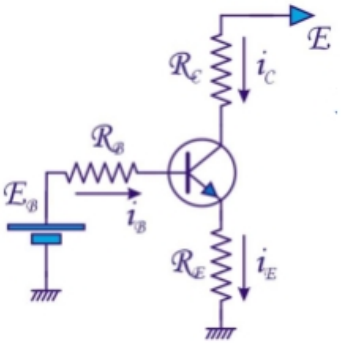


Figure 6.28

To make I_{CQ} independent from β, we should choose

$$R_B \ll \beta\, R_E \tag{6.21}$$

such that:

$$I_{CQ} \approx \frac{E_B - V_{BEQ}}{R_E} \tag{6.22}$$

In this equation, we suppose that E_B and R_E are constant and can be produced with high accuracy (there is still a temperature dependence, but this is relatively low). Suppose that V_{BEQ} can vary, which has an impact on I_{CQ}. This can be quantified with equation 6.22:

$$\Delta I_{CQ} = \frac{\Delta V_{BEQ}}{R_E} \tag{6.23}$$

A typical problem then goes as follows:

- The limits of V_{BEQ} are known, so we know ΔV_{BEQ}
- The limits of β are known: $(\beta_{min}, \beta_{max})$
- The value of I_{CQ} is given, with a certain precision ΔI_{CQ}

This problem can be solved by following these steps:

- With ΔV_{BEQ} and ΔI_{CQ}, determine $R_E = \frac{\Delta V_{BEQ}}{\Delta I_{CQ}}$,
- With R_E and the equation of the left loop, determine $E_B = R_E\, I_{CQ} + V_{BEQ}$,
- With β_{min}, choose a R_B such that $R_B \approx \beta_{min} R_E/10$. This guarantees that condition 6.21 is satisfied.
- With R_B and E_B, determine R_1 and R_2 based on the Thevenin equations.

This procedure allows us to determine R_1, R_2 and R_E. The only remaining unknown is R_C. We will determine this resistor by requiring that the operating point Q lies in the middle of the normal operating region. Refer to figure 6.27 of the biasing circuit with bypass capacitor capacitor. We assume that $C_E \to \infty$, such that it is a short circuit for small signals[3]. In this way, the static and dynamic load lines correspond to those of figure 6.27(b).
The slope of the dynamic load line is given by R_C. To determine this resistor, we need to set two points of the load line. We choose to limit the current I_C between 0 and $2I_{CQ}$. This corresponds to a maximum attainable current swing. The minimum voltage is $V_{CE,Sat}$, because below this value the transistor goes into saturation. The maximum voltage is $E - R_E\, I_{CQ}$, as can be derived from the figure. Note that in AC, the emitter is grounded, so the voltage at the emitter is fixed. Hence we compute R_E as the slope between these two points:

$$R_C = \frac{E - R_E\, I_{CQ} - V_{CE,Sat}}{2I_{CQ}} \tag{6.24}$$

This reasoning is graphically represented in figure 6.29.

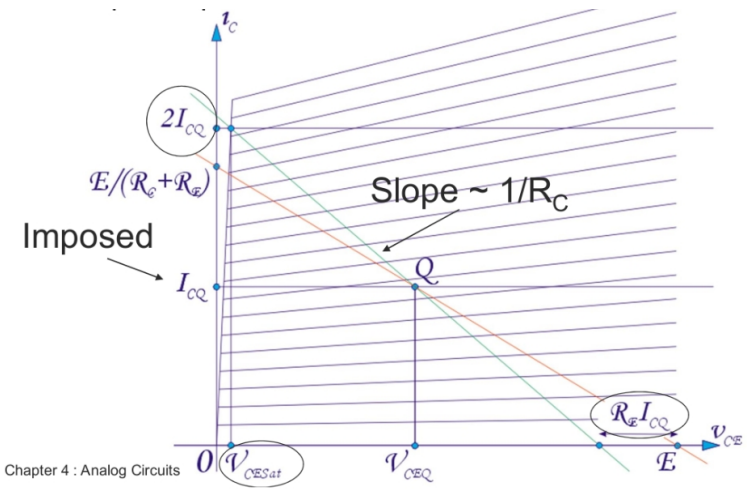


Figure 6.29: Determining R_C to place in Q in the middle of normal operating region

6.4.2 MOSFET Biasing

The goal of MOSFET biasing is to choose R_1, R_2 and R_S to reduce the impact of variations on V_T and K. Furthermore, R_D will be chosen such that the operating point Q lies in the middle of the normal operating region.

[3]Or we assume that the only frequencies of interest are much higher that $\omega_0/2\pi$.

The circuit we consider is the one in figure 6.30. As always, we replace the maze on the left by the Thevenin equivalent with E_G and R_G. We don't know the exact position of the i_{DS} vs v_{GS} curve because:

- The required value of I_{DSQ} is only given within certain limits: $I_{DSQ,min} < I_{DSQ} < I_{DSQ,max}$
- The manufacturer gives the value of $K = \mu C_{ox}$ within limits: $K_{min} < K < K_{max}$
- The same goes for the threshold voltage V_T: $V_{Tmin} < V_T < V_{Tmax}$.

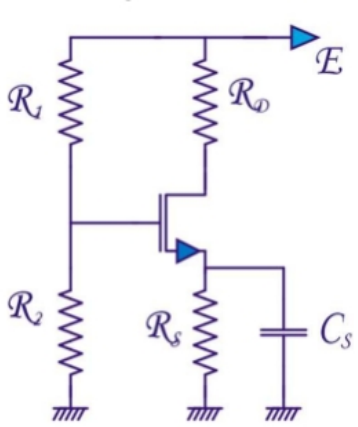


Figure 6.30

This allows us to draw a minimum (based on V_{Tmax} and K_{min}) and maximum curve (based on V_{Tmin} and K_{max}), as in figure 6.31. The intersection of these curves with resp. $I_{DSQ,min}$ and $I_{DSQ,max}$ gives two points on which the load line $E_G = v_{GS} + R_S i_{DS}$. This is the only way to ensure that the intersection between load line and the real curve gives a current between $I_{DSQ,min}$ and $I_{DSQ,max}$. The slope of the line between these two points determines thus R_S, while the intersection with the x-axis sets E_G.
As there is no DC current through R_G, its value doesn't really matter for setting the operating point. We can choose R_G freely and have an additional degree of freedom to determine R_1 and R_2 from the Thevenin equations[4].

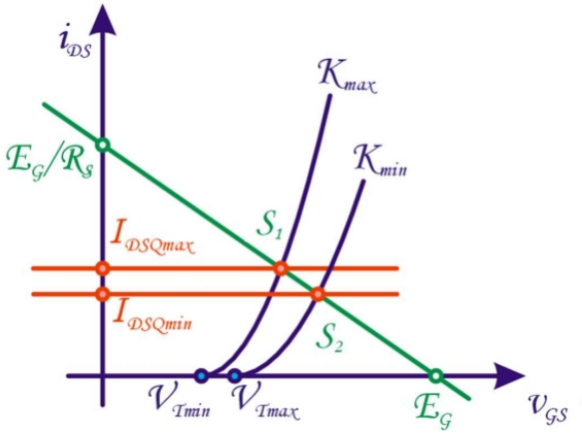


Figure 6.31

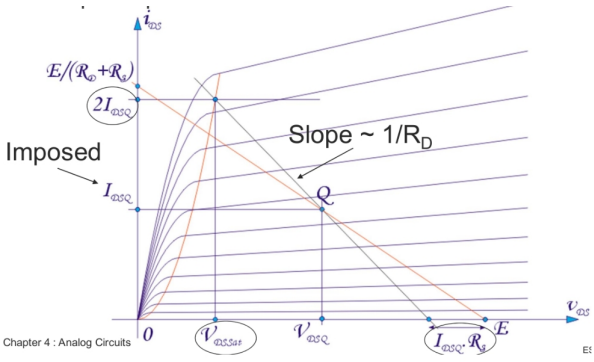


Figure 6.32

To determine R_D, we apply the same reasoning as for the BJT: we want to place Q in the middle of the saturation region. The minimum and maximum current is 0 and 2 I_{DSQ}; the corresponding voltage range is $V_{DS,Sat} < V_{DS} < E - R_S\ I_{DSQ}$. The value of $V_{DS,Sat} = V_{GS} - V_T$, the minimum V_{DS} to stay in saturation, has to be determined by solving $2I_{DSQ} =$

[4]In the exercises, R_G will be given.

$\frac{K}{2}\frac{W}{L}V_{DS,Sat}^2$ because at the edge of saturation, the required current is $2I_{DSQ}$:

$$V_{DS,Sat} = \sqrt{\frac{2I_{DSQ}}{\frac{K}{2}\frac{W}{L}}}$$

We then determine R_D as:

$$R_D = \frac{E - R_S\ I_{DSQ} - V_{DS,Sat}}{2\ I_{DSQ}} \tag{6.25}$$

6.5 The Small-Signal Model

Basically, the small-signal model of a (non-linear) component is a representation of the component that can be used as a substitute when we consider small signals. It should only contain linear elements (resistors, inductors, capacitors, linearly depended current or voltage sources, ...) because it is obtained by linearizing the behavior of the component around an operating point Q.
In section 6.2, we established the small-signal model for the diode. This was a resistance ρ_d in parallel with a capacitor C_j, as shown in figure 6.20. The values of both elements are set by the operating point (V_{DQ}, I_{DQ}). In this section, we will develop the small-signal model for a bipolar junction transistor and for a MOSFET.

6.5.1 BJT Small-Signal Model

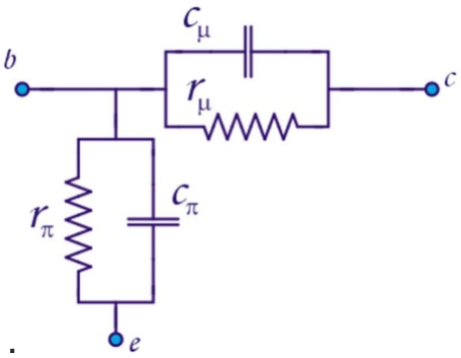


Figure 6.33: Just the pn-junctions

A bipolar junction transistor is nothing else than two pn-junctions put against each other. So we could just put the small-signal model of diode between base and emitter and between base and collector, as in figure 6.33 (for a npn transistor). However, in doing so, we don't have any component that mimics the transistor action, namely the dependence of i_c on i_b. To do this, we define a current gain h_{fe}

$$h_{fe} = \frac{di_C}{di_B} = \frac{i_c}{i_b} = \frac{d\beta i_b}{di_b} = \beta + \frac{d\beta}{di_b} i_b \approx \beta \tag{6.26}$$

and place a dependent current source $h_{fe}i_b$ between collector and emitter. Furthermore, since the output current between collector and emitter also depends on v_{ce} because of the

Early effect, we add an "Early" resistor r_c between both terminals. Figure 6.34 gives the entire small-signal model for a npn BJT. The different parameters depend on the biasing conditions:

- Resistance r_π is the diode resistance between base-emitter junction, thus: $r_\pi = \frac{v_{th}}{I_{BQ}}$.
- The base-collector junction is reversed biased, so $r_\mu \approx 0$.
- The Early resistance depends on the Early voltage V_E, which is about 40V: $r_c \approx \frac{V_E}{I_{CQ}}$.
- Often, we express i_c as function of v_{be}. The ratio between both is the *transconductance* g:

$$g = \frac{i_c}{v_{be}} = \frac{i_c}{r_\pi i_b} = \frac{h_{fe}}{r_\pi} \approx \frac{\beta I_{BQ}}{v_{th}} = \frac{I_{CQ}}{v_{th}}$$

By introducing this transconductance, we replace the current-dependent current source $h_{fe}\, i_b$ with a voltage-dependent current source $g\, v_{be}$. This is much better, because we can control I_{CQ} by choosing R_E and we can thus set g with high precision. This is not the case for β, as explained previously. The result is shown in the model in figure 6.35, which is also called *Giacoletto's model*.

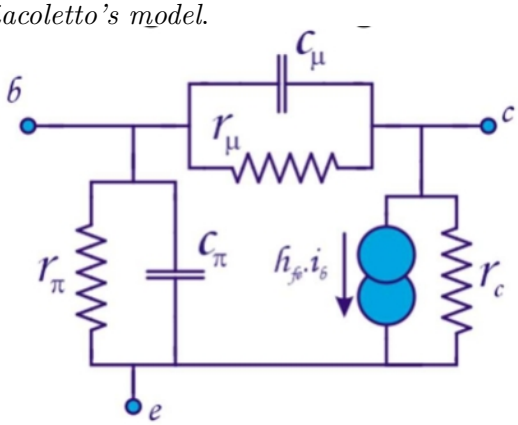


Figure 6.34: BJT small signal model

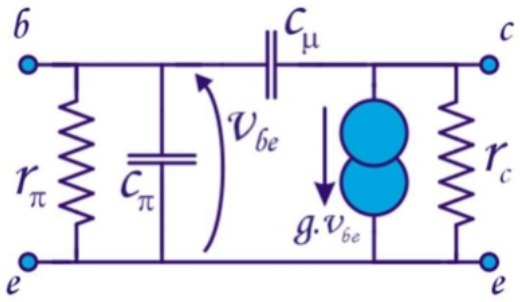


Figure 6.35: Giacoletto's model

Note that $C_\pi \gg C_\mu$ because the width of the depletion zone is much smaller between emitter and base than between base and collector (the latter is reversed biased while the former is forward biased) and $C \sim \epsilon/t$ with t the thickness of the depletion region. When working at low frequencies, the capacitors are omitted and replaced by open circuits to obtain the low-frequency model.

6.5.2 MOSFET Small-Signal Model

For the MOSFET transistor, we observe that:

- A voltage v_{gs} causes a current between drain and source. We model this by a transconductance g_m.
- Because of channel-length modulation, v_{ds} also has an impact on the current between drain and source. We model this with a resistor r_{ds}.
- The connections between gate and between source and gate and drain are capacitors: C_{gs} and C_{gd}.

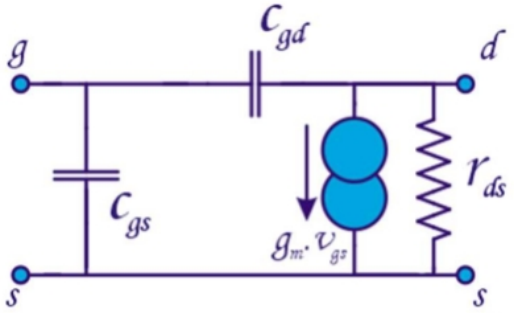


Figure 6.36: MOSFET Small signal model

All these elements are represented in figure 6.36. We can compute g_m, based on the $i_{DS} - v_{GS}$ characteristic (when $v_{GS} > V_T$):

$$\begin{aligned} i_{DS} &= \frac{K}{2}\frac{W}{L}(v_{GS} - V_T)^2 \\ \Rightarrow \frac{di_{DS}}{dv_{GS}} &= K\frac{W}{L}(v_{GS} - V_T) = \frac{2i_{DS}}{v_{GS} - V_T} \\ \Rightarrow g_m &= \frac{di_{DS}}{dv_{GS}} = \frac{i_{ds}}{v_{gs}} = \frac{2I_{DSQ}}{V_{DSQ} - V_T} \end{aligned} \tag{6.27}$$

As for r_{ds}, this quantity is related to the channel-length modulation factor λ from equation 5.6:

$$i_{DS} = \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(v_{GS} - V_T)^2\,(1 + \lambda v_{DS}) \tag{6.28}$$

This allows us to compute the change in i_{DS} for small variations of v_{DS}:

$$\begin{aligned} \frac{\partial i_{DS}}{\partial v_{DS}} &= \frac{1}{2}\mu_n C_{ox}\frac{W}{L}(v_{GS} - V_T)^2\,\lambda \\ &\approx \lambda\, I_{DSQ} \end{aligned} \tag{6.29}$$

This expression is the conductivity g_{ds} and thus $r_{ds} = \frac{1}{g_{ds}} \approx \frac{1}{\lambda\, I_{DSQ}}$.

6.5.3 Orders of magnitude

To estimate the values of the transistor parameters, we will assume that (a) a good value for the Early voltage ≈ 40 V and that (b) the designer should choose a small $V_{DS,Sat}$ to maximize g_m, typically ≈ 200 mV.
For the bipolar transistor, we observe that:

- $g_\pi = \frac{1}{r_\pi} = \frac{I_{BQ}}{v_{th}} = \frac{I_{CQ}}{\beta v_{th}} = \frac{g}{\beta}$
- $g_c = \frac{1}{r_c} = \frac{I_{CQ}}{V_E} = \frac{v_{th}}{V_E}\frac{I_{CQ}}{v_{th}} \approx \frac{g}{1600}$

and thus: $g \gg g_\pi \gg g_c$.
For the MOSFET, we find that:

- $g_{ds} = \frac{1}{r_{ds}} = \frac{I_{DSQ}}{V_E} = \frac{v_{DS,Sat}}{2V_E}\frac{2I_{DSQ}}{v_{DS,Sat}} \approx \frac{g_m}{400}$

and thus: $g_m \gg g_{ds}$.

Chapter 7

Amplifiers

In this section, we will study amplifiers: circuits that take a signal as input and produce an identical but magnified version at the output. We'll see the basic amplifier and the more stable four-resistor version. Next, we study the most common amplifier topologies - common emitter, common base and common collector - and what are their advantages and drawbacks. Finally, we study the differential amplifier and the operational amplifier or OPAMP.

7.1 Basic Amplifier

In this section, we will develop and improve a basic amplifier. The elementary circuit we will be using is shown in figure 7.1. Bias currents I_{BQ} and I_{CQ} are generated by the DC voltage source E_B. The time-varying voltage source v_i is the input signal, and is applied at the base of the transistor[1]. The output is measured at the collector.
As v_i increases, so does i_B, just as in figure 6.18. If the transistor is biased in the normal operating region, then $i_C = \beta i_B$ will also increase. As we move along the load line with increasing i_C, the voltage drop along R_C increases and the output voltage v_o decreases. We want to compute the voltage gain A_v. But before we do that, we first see how the input part of the circuit in figure 7.1 can be constructed.

7.1.1 Coupling Capacitance

Let's compute the corresponding Thevenin equivalent of the circuit in figure 7.2. To find e_0, apply Millman's theorem:

$$\begin{aligned} e_0 &= \frac{E/R_1 + e_i\, j\omega C_B}{1/R_1 + 1/R_2 + j\omega C_B} \\ &= \frac{R_2 E + e_i\, j\omega C_B R_2}{R_1 + R_2 + j\omega C_B R_1 R_2} \\ &= E\frac{R_2}{R_1+R_2}\frac{1}{1+j\omega C_B R_B} + e_i \frac{j\omega C_B R_B}{1+j\omega C_B R_B} \end{aligned} \tag{7.1}$$

with $R_B = \frac{R_1 R_2}{R_1+R_2}$. If ω is much smaller than the critical pulsation $\omega_c = \frac{1}{R_B C_B}$, then $e_0 \approx E\frac{R_2}{R_1+R_2}$. If $\omega \gg \omega_c$, then $e_0 \approx e_i$. The latter condition corresponds to assuming C_B is a

[1]Note that in reality this is the result of applying Thevenin's theorem to resistors R_1 and R_2

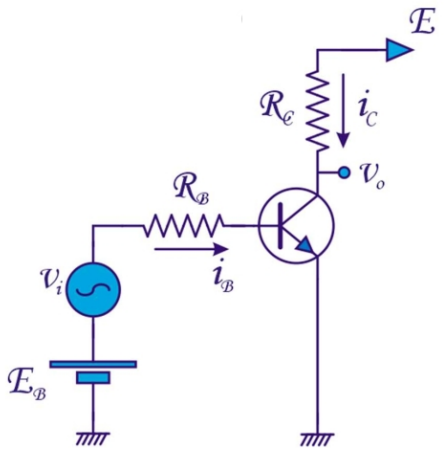


Figure 7.1: Simple amplifier

short-circuit. Another way to find equation 7.1, is to use the superposition principle: first, consider only E with $e_i = 0$, then consider e_i with $E = 0$, and add both results.
The circuit is thus equivalent to a DC source $E_B = E\frac{R_2}{R_1+R_2}$ in series with a small-signal, high-frequency source e_i, just as in figure 7.3. This means we can use this circuit to couple e_i to the input of the amplifier, while keeping the DC biasing, just as in figure 7.1. Capacitor C_B is a *coupling capacitor* because it "couples" v_i into the circuit.

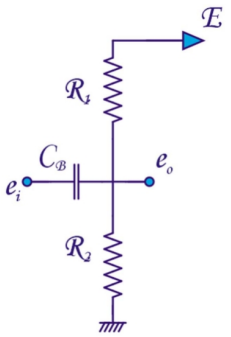


Figure 7.2

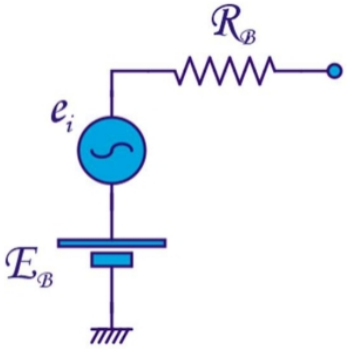


Figure 7.3

Voltage gain A_v

To calculate A_v, we draw the small-signal equivalent circuit, by replacing the npn-transistor by its small-signal model, and by grounding the DC voltage source E. We also assume that the frequency we consider is higher than $\frac{1}{R_B C_B}$, so we can consider C_B as a short-circuit. The small-signal circuit is shown in figure 7.4. The parameters of the model are set by the operating currents and voltages:

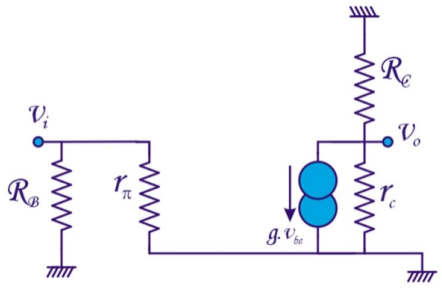


Figure 7.4: Amplifier - small-signal model

- $r_\pi = \frac{v_{th}}{I_{BQ}}$,
- $r_c = \frac{V_E}{I_{CQ}}$,
- $g = \frac{I_{CQ}}{v_{th}}$

In the equivalent circuit, we apply Millman in v_o:

$$v_o = \frac{-g\ v_{be}}{g_c + G_C}$$

and we see that $v_{be} = v_i$. As a consequence:

$$A_v = \frac{v_o}{v_i} = -g\ (r_c||R_C) \tag{7.2}$$

For a MOSFET, we would have found a similar expression: $A_v = -g_m(r_{ds}||R_D)$. Note that the gain is negative, because an increase in i_b and thus in i_c will lead to a larger drop across R_C and will decrease v_o, as explained previously.
The gain can be increased by:

1. Increasing g by setting a higher I_{CQ}. This will also decrease r_c, but this is usually not a problem since most of the times $r_c \gg R_C$ and hence $R_C||r_c \approx R_C$. In that case, $A_v \approx -g\ R_C$.
2. Increase R_C. The drawback is that this decrease the potential swing of v_o.

The maximum voltage gain we can obtain for a BJT is found when $R_C \to \infty$. Then is

$$A_{v,max} = -gr_c = -\frac{I_{CQ}}{v_{th}}\frac{V_E}{I_{CQ}} \approx -40 \times \frac{1}{0.026} \approx -1600$$

For a MOSFET, $A_{v,max}$ is about -400.

7.1.2 The 4-resistor amplifier

We will study a more general circuit, namely the amplifier with 4 biasing resistors that we saw before and is reproduced in figure 7.5(left). As before, we assume that the input frequency of interest is such that we can consider C_B as a short circuit. Note that the small-signal circuit would be the same for an n-channel MOSFET, if $r_\pi \to \infty$.
To compute A_v, apply Millman at both the emitter and collector (output) node:

- At collector: $v_o = \frac{g_c v_e - g(v_i - v_e)}{G_c + g_c}$ because $v_{be} = v_i - v_e$. Note that we used conductivities. For example, $G_c = 1/R_c$.
- At emitter: $v_e = \frac{g_\pi v_i g_c v_o + g(v_i - v_e)}{G_E + g_c + g_\pi}$

After eliminating v_e, we obtain:

$$\begin{aligned} A_v &= \frac{-gR_C r_c r_\pi + R_C R_E}{(r_c + R_C)(R_E + r_\pi) + r_\pi R_E(1 + gr_c)} \\ &= -g\frac{r_c R_C}{r_c + R_C}\frac{r_\pi - \frac{R_E}{gr_C}}{R_E + r_\pi(1 + R_E\frac{1+gr_c}{r_c+R_C})} \\ &\approx -\frac{R_C}{R_E} \end{aligned} \tag{7.3}$$

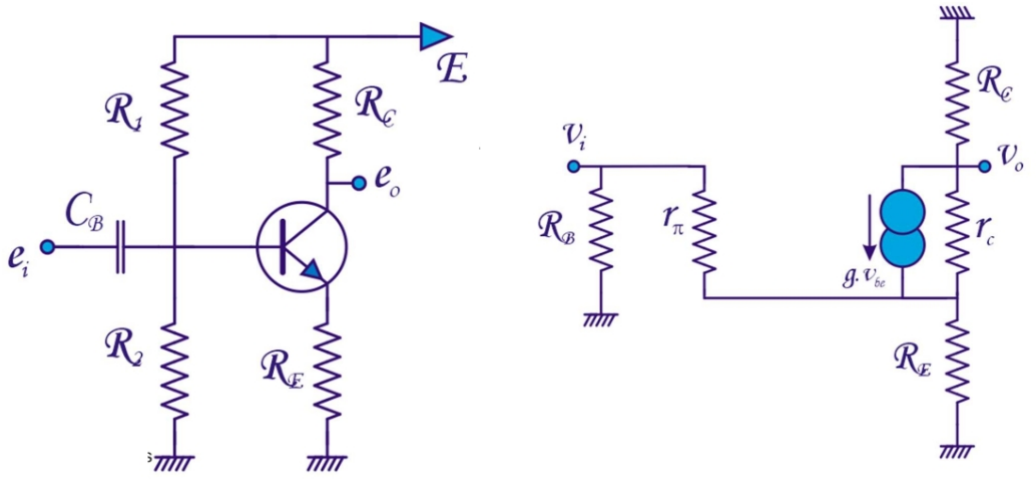


Figure 7.5: Four-resistor amplifier (left) and small-signal equivalent circuit (right)

where we assumed that $r_c \gg R_C$ and $gr_c \gg 1$.
It is logical that $A_v \approx -R_C/R_E$: in a first approximation, as the base voltage increases with v_i, the emitter voltage will follow because they are tied together by the drop across base-emitter junction, which is about 0.6 V. If the emitter voltage increases by v_i, the emitter current will increase by $\frac{v_i}{R_E}$. This is also approximately the increase in collector current, thus the voltage drop increase at R_C is equal to $v_o \approx -R_C \frac{v_i}{R_E}$.
We compare this result to the version with no emitter resistance, for which $A_v \approx -g\, R_C$. We take as nominal values $R_C = 1k\Omega$, $R_E = 500\Omega$ and $g = 40mA/V$. Without R_E, we find a gain of 40, while with R_E, the gain is 2. Thus, while R_E is necessary to obtain a stable bias point, its presence significantly reduces the gain. Hence we use a bypass capacitor as in figure 6.27.

Frequency analysis

We will study how the amplifier gain $A_v = \frac{v_o}{v_i}$ depends on frequency. To do this, we establish the small-signal circuit for the four-resistor amplifier with bypass capacitance C_E and coupling capacitance C_B, as in figure 7.6. The parallel combination of R_E and C_E gives:

$$Z_E = \frac{R_E}{1 + j\omega R_E C_E} \tag{7.4}$$

and we substitute this expression for R_E in equation 7.3. At the same this, we multiply by $\frac{j\omega R_B C_B}{1+j\omega R_B C_B}$, just as in equation 7.1.

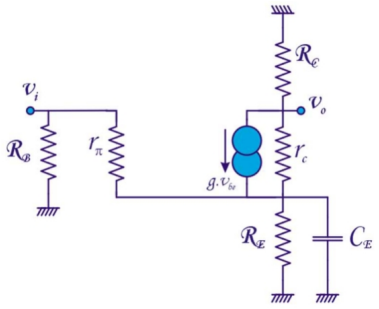


Figure 7.6

$$A_v = \frac{j\omega R_B C_B}{1 + j\omega R_B C_B} \frac{-gR_C r_c r_\pi + R_C Z_E}{(r_c + R_C)(Z_E + r_\pi) + r_\pi Z_E (1 + gr_c)} \tag{7.5}$$

If we set $T_B = 1/(R_B C_B)$ and $T_E = 1/(R_E C_E)$, this equation can be simplified to:

$$A_v \approx -A_{vE} \frac{j\omega T_B}{1 + j\omega T_B} \frac{1 + j\omega T_E}{1 + j\omega T_E \frac{A_{vE}}{A_{v0}}} \tag{7.6}$$

with $A_{vE} = \frac{R_C}{R_E}$ and $A_{v0} = g\frac{R_C r_c}{R_C + r_c}$. This transmittance has two poles in $\omega = 1/T_B$ and $\omega = \frac{A_{v0}}{T_E\ A_{vE}}$ and zeros in $\omega = 0$ and $\omega = 1/T_E$. With $A_{v0} \gg A_{vE}$ and a correct choice for $R_B C_B$ and $R_E C_E$, we have:

$$1/T_B < 1/T_E < \frac{A_{v0}}{T_E\ A_{vE}}$$

Figure 7.7 shows the bode plot of $A_v(\omega)$. We find four different domains based on the

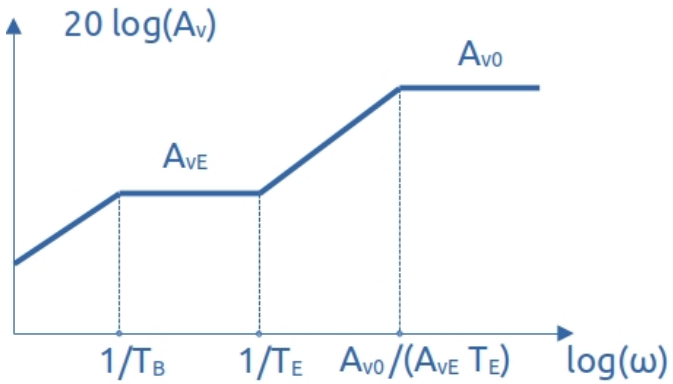


Figure 7.7: $4R\ A_v$ Bode Curve

frequency of the input signal:

1. $\omega < 1/T_B$: the signal v_i is not yet coupled into the base of the amplifier through capacitance C_B.
2. $1/T_B < \omega < 1/T_E$: capacitance C_B can be considered as a short circuit. However, ω is to small to short-circuit C_E and bypass R_E. The gain is thus about $-R_C/R_E$.
3. $1/T_E < \omega < A_{v0}/(A_{vE}T_E)$: as ω increases, R_E starts to get bypassed.
4. $\omega > A_{v0}/(A_{vE}T_E)$: the maximum gain (with bypassed R_E) is reached: $A_{v0} = -g(R_C||r_c)$. This is the domain in which we want to use the amplifier.

It is important to note that the critical pulsation is not $1/T_E$ but rather $\omega_{crit} = \frac{A_{v0}}{A_{vE}T_E}$. Note also that $A_{vE} \approx -\frac{R_C}{R_E}$ and $A_{v0} \approx -g\ R_C$, so the critical pulsation is $\frac{1}{T_E}\frac{A_{v0}}{A_{vE}} \approx \frac{1}{R_E C_E}\frac{gR_C}{\frac{R_C}{R_E}} = \frac{g}{C_E}$.

7.2 Basic Topologies

7.2.1 Common Emitter Amplifier (CEA)

The amplifier configuration of the previous section is the *common-emitter* configuration: the input is at the transistor base (gate) and the output is at the collector (drain). For the frequencies of interest, the emitter (source) is bound to ground and thus at a "common" voltage. Other configurations are the common-base and common-collector. We will study these configurations in the next sections.
Our goal is to establish the voltage gain, and input- and output impedances for the common-emitter configuration. For this purpose, we once again draw the small-signal equivalent circuit as in figure 7.8.

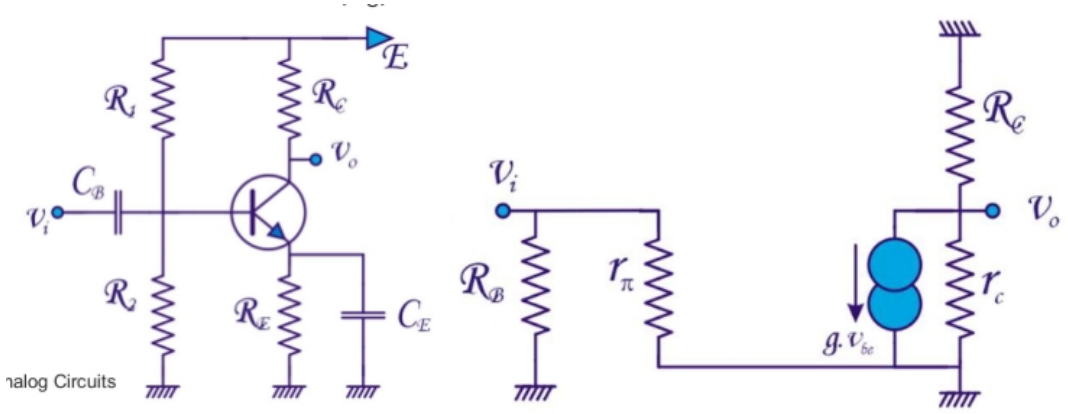


Figure 7.8: CEA: circuit (left) and small-signal equivalent (right)

- Voltage gain A_v: For an input voltage v_i, $v_{be} = v_i$. Thus $v_o = -gv_i(R_C||r_c)$ and $A_v = -g(R_C||r_c) = -\frac{g}{g_c+G_C}$ as we've seen before.
- Input impedance $Z_i = v_i/i_i = \frac{1}{g_\pi+G_B} = r_\pi||R_B$.
- Output impedance: to compute Z_o:
  - Shorten v_i to ground,
  - Apply v_o (or i_o) to the output,
  - Compute i_o (or v_o),
  - The output impedance $Z_o = \frac{v_o}{i_o}$.

  The effect of shortening the input to ground is that $v_{be} = 0$ is in figure 7.8. Thus the current source with transconductance g can be omitted. We see that then the output impedance is the parallel combination of R_C and r_c: $Z_o = \frac{1}{G_C+g_c}$.

7.2.2 Common Base Amplifier (CBA)

In a common-base configuration, the base of the transistor is kept at a constant voltage (i.e. an AC ground). The input signal is applied to the emitter and the output voltage is measured

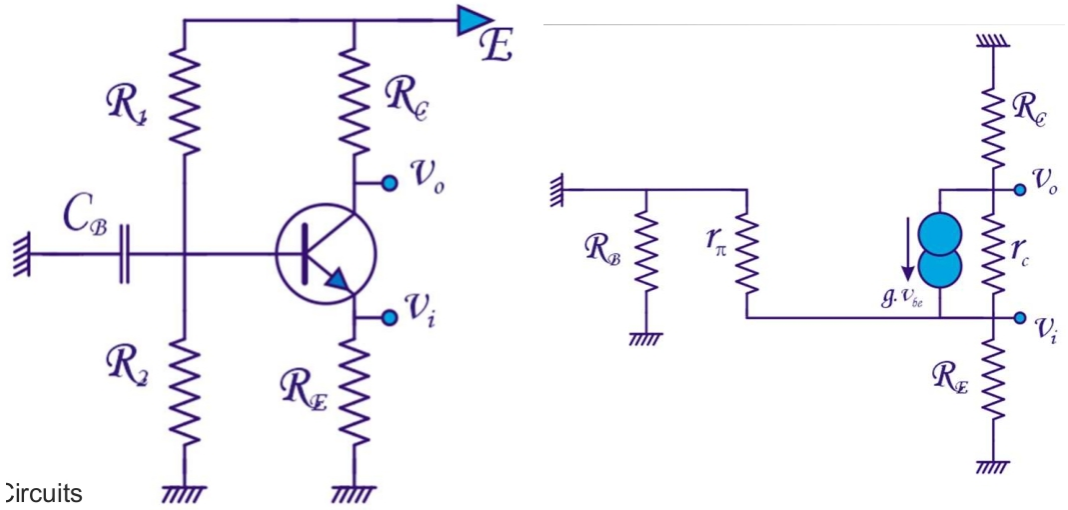


Figure 7.9: CBA: circuit (left) and small-signal equivalent (right)

at the collector. The circuit is shown in the left part of figure 7.9, with the small-signal circuit on the right. Obviously, a bypass capacitance is not added.

- Voltage gain A_v: we see that $v_{be} = -v_i$. In the output node, we can write:

$$\begin{aligned} v_o &= \frac{g_c v_i - g(-v_i)}{G_C + g_c} \\ \rightarrow A_v &= \frac{v_o}{v_i} = \frac{g + g_c}{G_C + g_c} \\ &\approx \frac{g}{G_C + g_c} = g(R_C || r_c) \end{aligned} \tag{7.7}$$

  because $g \gg g_c$. Notice how the gain is positive, and the same (in absolute value) as for the common-emitter configuration.

- Input impedance Z_i: We compute the current drawn at the input node:

$$i_i = G_E v_i + g_\pi v_i + g_c(v_i - v_o) - g v_i \tag{7.8}$$

  Substituting the expression for v_o: $v_o = \frac{g+g_c}{G_C+g_c} v_i$ into this equation gives:

$$\begin{aligned} Z_i &= \frac{g_c + G_c}{(g + g_c)G_C + (g_c + G_C)(g_\pi + G_E)} \\ &\approx \frac{1}{G_E + g} \end{aligned} \tag{7.9}$$

  This last expression is the parallel combination of R_E with a resistance $\frac{1}{g}$: if we look in the emitter (or source), we see an impedance $1/g$ (or $1/g_m$).

- Output impedance Z_o: by shorting the input, $v_{be} = 0$ thus there is no current through the transconductance. We see R_C in parallel with r_c:

$$\Rightarrow Z_o = \frac{1}{G_C + g_c},$$

  just as for the CEA.

7.2.3 Common Collector Amplifier (CCA)

In a common collector amplifier, the input is applied to the base, and the output is measured at the emitter, as in figure 7.10. The common node - the AC ground - is the collector, thus no collector resistance R_C is needed. Furthermore, we don't use a decoupling capacitor C_E. In a first approximation, we can say that $v_I - v_O = v_{BE} \approx 0.6$ V and remains constant. That why $\frac{v_o}{v_i} \approx 1$ and we also call this topology a *follower* because the output follows the input.

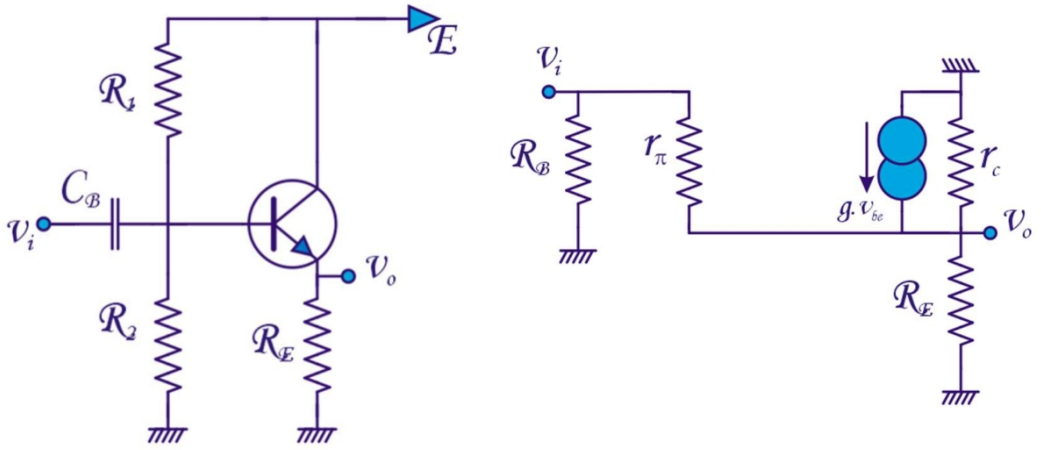


Figure 7.10: CCA: circuit (left) and small-signal equivalent (right)

- Voltage gain A_v: we see that $v_{be} = v_i - v_o$. In the output node, we can write:

$$\begin{aligned} v_o &= \frac{-g\ v_{be} + g_\pi v_i}{g_\pi + g_c + G_E} \\ &= \frac{-g\ (v_o - v_i) + g_\pi v_i}{g_\pi + g_c + G_E} \\ \rightarrow (g + g_\pi + g_c + G_E)\ v_o &= (g + g_\pi)\ v_i \\ \rightarrow A_v = \frac{v_o}{v_i} &= \frac{g + g_\pi}{g + g_\pi + g_c + G_E} \approx \frac{g}{g + G_E} \approx 1 \end{aligned} \tag{7.10}$$

- Input impedance Z_i.
  Consider the circuit initially without R_B. Then

$$\begin{aligned} i_o &= g_\pi(v_i - v_o) \\ &= g_\pi(1 - \frac{g + g_\pi}{g + g_\pi + g_c + G_E})v_i \\ &= g_\pi(\frac{g_c + G_E}{g + g_\pi + g_c + G_E})v_i \\ &\approx g_\pi \frac{G_E}{g} \\ \rightarrow Z_i &= R_E \frac{g}{g_\pi} = \beta R_E \end{aligned} \tag{7.11}$$

  and thus $Z_i = \beta R_E \,||\, R_B$.

- Output impedance Z_o:

$$i_o = (G_E + g_c + g_\pi + g)\, v_o$$

and thus

$$Z_o = \frac{1}{G_E + g_c + g_\pi + g} \approx \frac{1}{g + G_E} \approx \frac{1}{g}$$

7.2.4 Comparison of Topologies

The characteristics of the different topologies - both for BJT as for the MOSFET (Common source, gate, and drain configurations) - are summarized in table 7.2.4. We only use intrinsic parameters of the transistors.

	Z_i	Z_o	$\lvert A_v \rvert$
CEA	r_π	r_c	$g\, r_c$
CBA	$1/g$	r_c	$g\, r_c$
CCA	βR_E	$1/g$	1
CSA	∞	r_{ds}	$g_m r_{ds}$
CGA	$1/g_m$	r_{ds}	$g_m r_{ds}$
CDA	∞	$1/g_m$	1

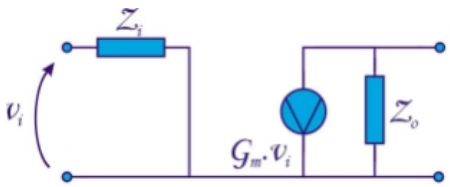


Figure 7.11

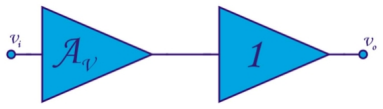


Figure 7.12

Figure 7.11, the schematic representation of an amplifier as seen in chapter 1, shows that as a general rule $|A_v| = g\, Z_o$. This can be verified in table 7.2.4. From this figure, we can also deduce that a good amplifier needs (a) a high input impedance to avoid drawing a large current from the previous stage and (b) a low output impedance to avoid making the output voltage depended on the impedance of the next stage. The only suitable configuration is the common collector (or CDA), but this amplifier has a gain of ≈ 1. In conclusion, to implement a good amplifier we need a cascade of:

- An amplifier with high gain, medium Z_i and high Z_o,
- A buffer stage with gain ≈ 1, high Z_i and low Z_o, as in figure 7.12.

Consequently, the gain will be high, and in- and output impedances will be as required.

7.3 Differential Amplifier

7.3.1 Definition

Until now, the amplifiers we studied only had a single input terminal and a single output terminal. We would like to create an amplifier that amplifies the *difference* between two voltages. This can be useful to compare two signals as in an operational amplifier, or to remove the noise on two signals when this noise is significantly correlated (e.g. because it was created by the same noise source). There are 2 inputs v_p and v_n (the positive and negative

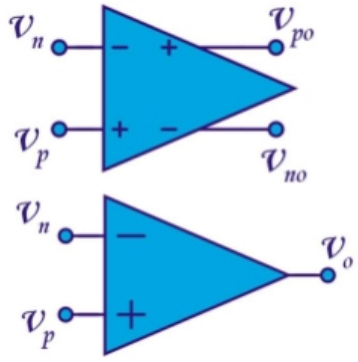


Figure 7.13: Differential amplifier with two (top) or one (bottom) output.

terminal). The goal is to amplify the difference $v_p - v_n$. This can be done with one or two outputs. In the latter case - also called the *differential* case - the voltage between outputs v_{po} and v_{no} is:

$$v_{os} = v_{po} - v_{no} = A_v\ (v_p - v_n)$$

as in the top of figure 7.13. The other option is the single-output, asymmetrical amplifier:

$$v_o = A_v\ (v_p - v_n)$$

To make the analysis easier, we split the signal in two components:

- A *differential mode*: $v_d = v_p - v_n$,
- A *common mode* $v_c = \frac{v_p + v_n}{2}$

This means that we can write $v_p = v_c + \frac{v_d}{2}$ and $v_n = v_c - \frac{v_d}{2}$. In a typical scenario, the common-mode signal can be a lot larger (order of magnitude several volts) than the differential mode (several milivolts).
With a single output, the output voltage is thus $v_o = A_d v_d + A_c v_c$:

- The differential gain A_d, which is actually what we want, so it has to be as high as possible.
- The common mode gain A_c, which is a gain that we want to keep as low as possible - we want to reject the common mode. Ideally, it should be zero, but we will see that this is not possible (at least for a single-output amplifier).

We also define two so-called rejection ratios: the common-mode rejection ratio (CMRR) which is the ratio between A_d and A_c: CMRR $= A_d/A_r$, and the power supply rejection ration PSRR, which expresses how variations in the power supply have an impact on the output. Both rejection ratios should be as high as possible, to remove unwanted, parasitic components in the output signal. They are typically expressed in decibel and in a good amplifier, are about 100 - 120 dB.

7.3.2 Implementation

The circuit to accomplish the requirements from the previous section, is shown in figure 7.14. We use two branches with identical npn transistors and identical collector resistances R. Both emitters are tied to a common node with voltage v_e. The bases of both transistors are the inputs. The outputs are measured at or between the collectors. The resistance R_E carries a current I_{RE}.

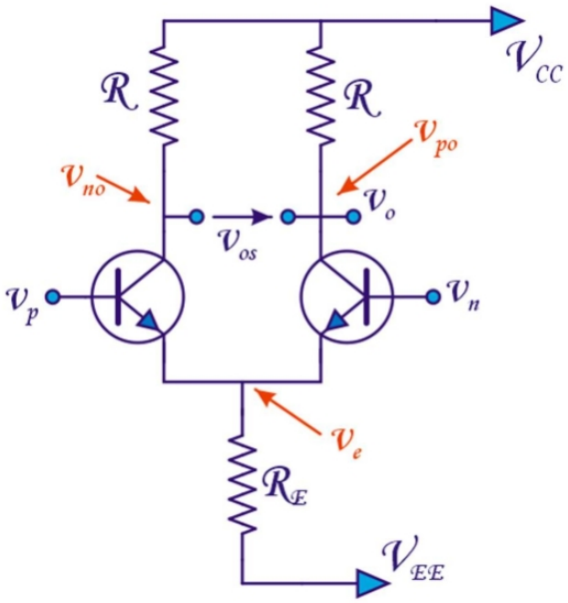


Figure 7.14: Structure of the differential amplifier

7.3.3 Common Mode

If the input nodes v_n and v_p are at the same voltage, $v_n = v_p = v_c$, the two branches carry an identical current $I_{RE}/2$ and the outputs v_{no} and v_{po} are both equal to $V_{CC} - R\ I_{RE}/2$. The differential output v_{os} is than zero. Because of the symmetry in the circuit, this output is very robust and depends only on the differential signals. However, if there is some mismatch between both branches, e.g. the transistors or resistances are not completely identical, this is no longer the case.
When $v_n = v_p = v_c$, then $v_e = v_c - V_{BEQ} \approx v_c - 0.6$ V. The current in both branches is equal to $i_c = \frac{I_{RE}}{2} = \frac{v_e - V_{EE}}{2R_E}$ and $v_{op} = v_{on} = v_o = V_{CC} - R\ i_c$. The symmetrical output $v_{os} = v_{po} - v_{po}$ is still zero, as it should be. The single output v_o however, varies with v_c.

This variation should be as small as possible.
To study the common-mode response, we can split the circuit into two parts. In a first stage, we split resistance R_E into two parallel resistors of value $2R_E$, as in figure 7.15. Because of the symmetry and because both inputs are the same, there can be no current in the wire connecting both emitters, so we remove it. So, in effect, we can study both halves independently which means we only have to analyze the circuit in figure 7.16.

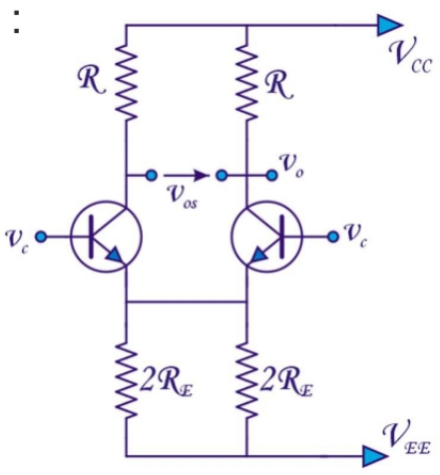


Figure 7.15

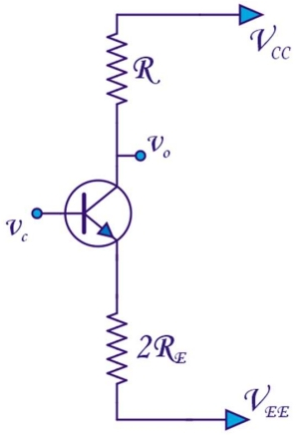


Figure 7.16

7.3.4 Differential Mode

With the common mode v_c is zero, $v_p = +v_d/2$ and $v_n = -v_d/2$. This is a purely differential signal: if v_p increases, v_n would decrease with the same amount. To demonstrate that in this case $v_e = 0$, we draw the small-signal equivalent circuit as in figure 7.17 and use Millman's theorem to compute the voltage in v_{on}, v_{op} and v_e. Note that v_{be} in the left branch is $\frac{v_d}{2} - v_e$, and $-\frac{v_d}{2} - v_e$ in the right branch.

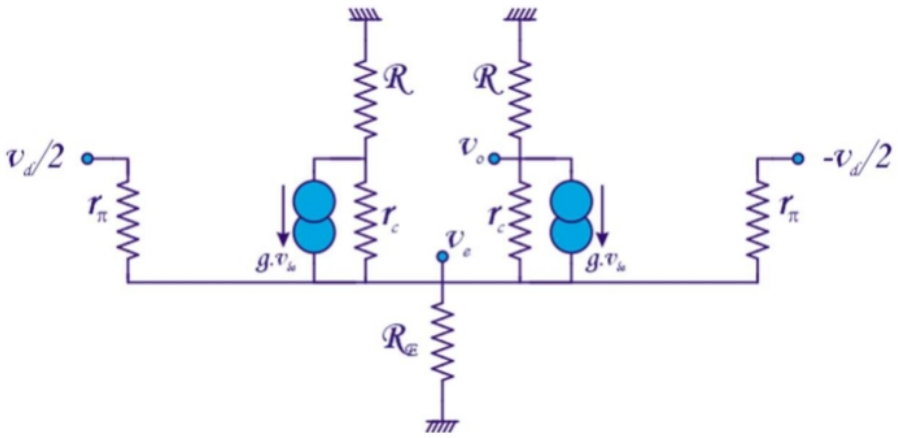


Figure 7.17: Differential mode: small-signal model

$$v_{po} = \frac{g_c v_e - g(-\frac{v_d}{2} - v_e)}{G + g_c}$$
$$v_{no} = \frac{g_c v_e - g(\frac{v_d}{2} - v_e)}{G + g_c}$$
$$v_e = \frac{g_c v_{po} + g_c v_{no} + g(-\frac{v_d}{2} - v_e) + g(\frac{v_d}{2} - v_e) - g_\pi \frac{v_d}{2} + g_\pi \frac{v_d}{2}}{G_E + 2g_c + 2g_\pi}$$

Substituting the expressions for v_{on} and v_{op} in the one for v_e gives:

$$v_e = \frac{g_c \frac{g_c v_e + g v_e}{g_c + G} + g_c \frac{g_c v_e + g v_e}{g_c + G} - g v_e - g v_e}{G_E + 2g_c + 2g_\pi}$$

Because of the symmetries, all mentions of v_d have disappeared in this expression. The solution is $v_e = 0$.
Building on the knowledge that $v_e = 0$ in a purely differential input signal (i.e. v_E doesn't change when both v_{op} and v_{on} change with equal magnitude but opposite sign), we can draw the following circuit for the differential input of figure 7.18. In this circuit, the current through resistance I_{RE} is constant because v_E is constant for differential input signals. The collector current I_{CQ} is constant and set by the common mode: $I_{CQ} = I_{RE}/2$.
Since we apply $+v_d/2$ to the left input, the current in the left loop increases by $i_c \approx g\ v_{be} = g\ v_d/2$ (remember: $v_e = 0$). The current in the right loop decreases by the same amount: $i_c \approx -g\ v_d/2$. So any additional current in the left loop flows in the right loop, keeping the current through R_E constant, as established previously. The output voltages change as well: $v_{on} = -R\ g\ v_d/2$ and $v_{op} = +R\ g\ v_d/2$ and thus $v_{os} = g\ R\ v_d$.

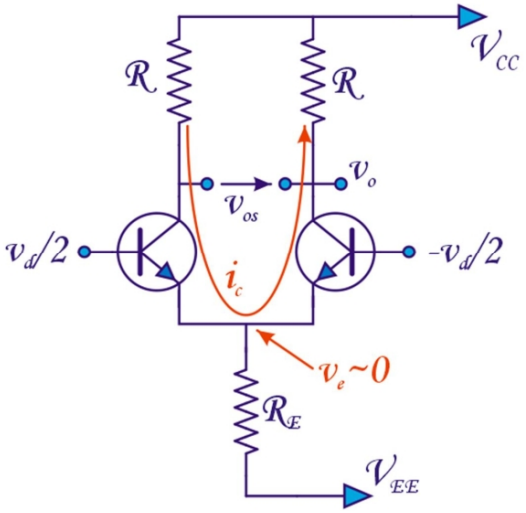


Figure 7.18

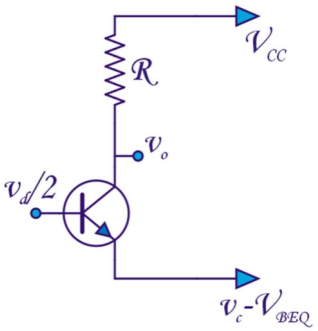


Figure 7.19

We can thus summarize that when we apply a differential signal, the emitter voltage v_E stays fixed and equal to $\approx v_c - 0.6$ V and the circuit can be studied by considering only one branch, namely the one in figure 7.19.
This means we have an equivalent circuit to study the common mode in figure 7.16 and a circuit to study the differential mode, in figure 7.19.

7.3.5 Load-line Analysis

The current i_C is determined by the common mode, from figure 7.16:

$$i_C = \frac{v_c - V_{BEQ} - V_{EE}}{2R_E} \tag{7.12}$$

The load line equation is given by:

$$V_{CC} - (v_c - V_{BEQ}) = Ri_C + v_{CE} \tag{7.13}$$

from figure 7.19.
The operating point Q is found by the intersection of these two lines, as in figure 7.20. Note how v_c impacts both lines: the current shifts up and down, and the load line moves parallel with varying v_c. This observation makes it very difficult to control the operating point. Commonly, the outer limits of v_c are given, so an estimate of the range of Q can be made.
As v_d changes, one transistor moves up along the load line, while the other one moves down because the load line equation was derived from the circuit where we supposed the input was symmetrical around v_c.

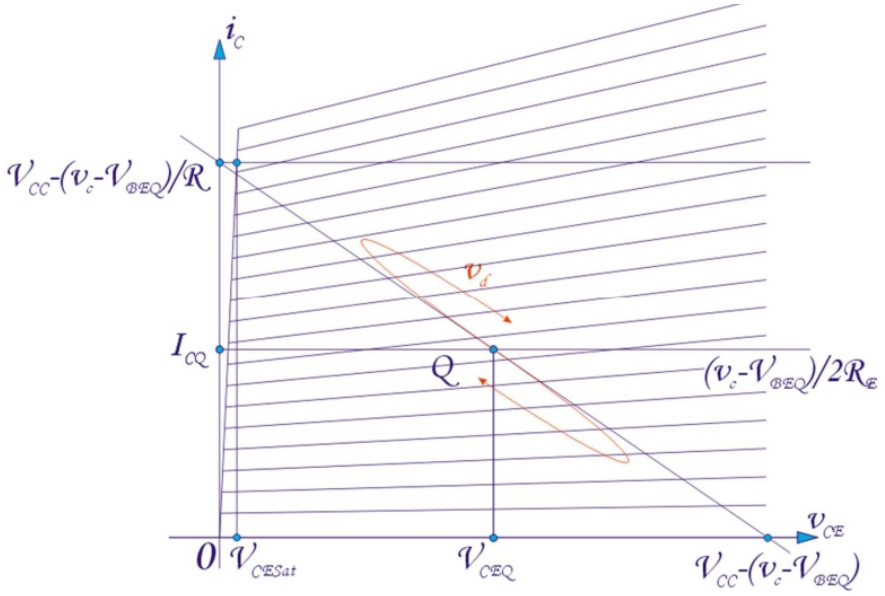


Figure 7.20

7.3.6 Common & Differential Gain

The *differential gain* A_d can easily be computed by transforming the circuit in figure 7.19 to its small-signal equivalent. The emitter voltage is an AC ground because this voltage doesn't change with v_d. The circuit is in fact a common-emitter amplifier. With r_c in parallel with R and $v_{be} = v_d/2$, we find:

$$A_d = \frac{v_o}{v_d} = -\frac{g(R||r_c)}{2} \tag{7.14}$$

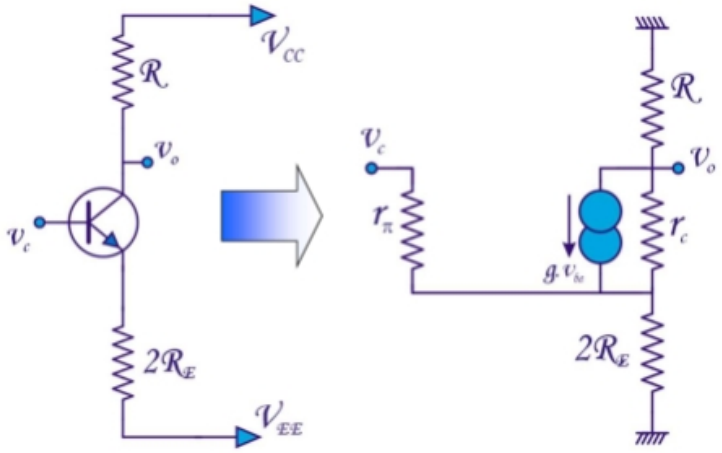


Figure 7.21

The *common-mode gain* A_c can be calculated by using the small-signal equivalent circuit for the circuit in figure 7.16. The result is found in figure 7.21.
This circuit is the same as the one in figure 7.5, but with $2R_E$ in stead of R_E. So we can reuse the result from equation 7.3:

$$\begin{aligned} A_c = \frac{v_o}{v_c} &= \frac{-gRr_cr_\pi + 2RR_E}{(r_c+R)(2R_E+r_\pi) + 2r_\pi R_E(1+gr_c)} \\ &\approx -\frac{R}{2R_E} \end{aligned} \tag{7.15}$$

for a single output. For the differential output, $A_c = 0$.
From the results in equations 7.14 and 7.15, we find that:

$$\begin{aligned} CMRR = \frac{A_d}{A_c} &\approx \frac{-\frac{gR}{2}}{-\frac{R}{2R_E}} \\ &\approx g\,R_E \end{aligned} \tag{7.16}$$

This means that is we want to increase the CMRR, we have to increase R_E. However, if we increase R_E, and since the voltage at the emitters is equal to $v_c - V_{BEQ}$, the current I_{RE} through R_E decreases. For each transistor, we have $I_{CQ} = I_{RE}/2$, and $g = \frac{I_{CQ}}{v_{th}}$. So each increase in R_E will lead to a decrease in I_{CQ} and thus also in a proportional decrease in g. Hence the CMRR will remain about the same.
A better way to increase R_E is by using the circuit in figure 7.22, where we add an additional transistor between R_E and v_E. We can explain the functioning as follows: with a fixed V_{BB}, the voltage at the emitter of the added transistor is $V_{BB} - V_{BEQ} = V_{BB} - 0.6$ V. This voltage is (relatively) fixed, and controls the current I_{CQ} :

$$I_{CQ} = \frac{1}{2}\frac{V_{BB} - V_{BEQ} - V_{EE}}{R_E}$$

This current is fixed and doesn't depend (in a first approximation) on the collector voltage of the transistor (as long as this voltage is high enough). As the collector voltage can vary and

the current remains constant, the collector sees a very large resistance.

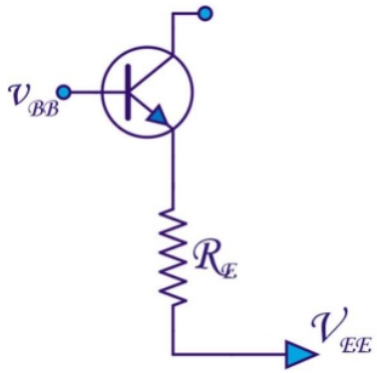


Figure 7.22

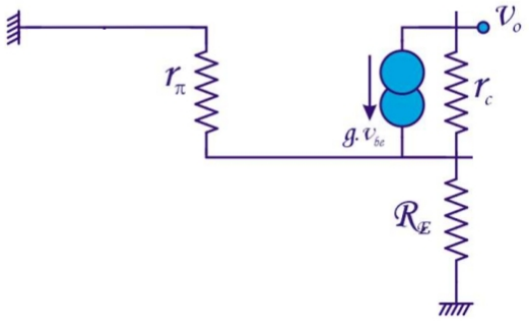


Figure 7.23

More formally, we can sketch the small-signal circuit as in figure 7.23 and compute the output impedance as seen in node v_o:

$$i_o = g_c(v_o - (-v_{be})) + gv_{be}$$

Since i_o also passes through the parallel combination of R_E and r_π, we can state that:

$$i_o = -(G_E + g_\pi)v_{be}$$

and thus:

$$\begin{aligned} i_o &= g_c v_o - (g + g_c)\frac{i_o}{G_E + g_\pi} \\ (G_E + g_\pi)i_o &= (G_E + g_\pi)g_c v_o - (g + g_c)i_o \\ (G_E + g_\pi + g + g_c)i_o &= (G_E + g_\pi)g_c v_o \\ \Rightarrow Z_o = \frac{v_o}{i_o} &= \frac{G_E + g_\pi + g + g_c}{(G_E + g_\pi)g_c} \\ &\approx \frac{g}{g_c G_E} \\ \Rightarrow Z_o &\approx (gr_c)\ R_E \end{aligned} \tag{7.17}$$

The resistance R_E is thus multiplied by the intrinsic transistor gain $gr_c \approx 1600$. The CMRR is thus $g(gr_c)R_E = g^2 r_c R_E$.

7.3.7 Power-Supply Rejection Ratio

To compute the power supply gain A_{cc}, we put the two inputs to AC ground an compute how v_o varies if the power supply undergoes a voltage change v_{cc}. Using the small-signal equivalent circuit of figure 7.19 with $v_d = 0$, we find that $v_o = \frac{r_c}{r_c + R_C} v_{cc} \approx v_{cc}$ so the supply voltage variation is almost completely transferred to the output: $A_{cc} \approx 1$. The power supply rejection ratio PSRR is thus equal to $\left|\frac{A_d}{A_{cc}}\right| \approx |A_d| \approx \frac{gR}{2}$.

7.4 Operational Amplifier

An operational amplifier or *OPAMP* is basically a differential amplifier with very high gain. It is typically constructed with a differential amplifier as input stage and an additional common-emitter stage to boost the gain of the differential amplifier. Schematically, we can see the internal structure of an OPAMP as in figure 7.24 where there are 4 stages:

1. A single-output differential amplifier as input stage,
2. A buffer stage to avoid that the next stage loads the output of the differential amplifier. This buffer stage has a high input impedance and low output impedance and a gain $A_v \approx 1$, like a common-collector amplifier.
3. A high-gain stage, like a common-emitter amplifier,
4. Another buffer stage to isolate the load circuitry from the OPAMP.

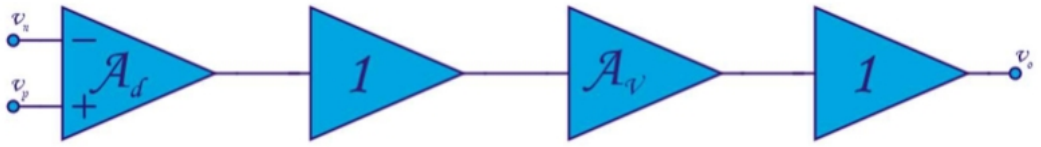


Figure 7.24

An OPAMP is very close to an ideal amplifier, which we will see in the next section. After that, we study the real OPAMP and see how non-idealities will impact its behavior.

7.4.1 The Ideal Amplifier

An OPAMP is represented by the symbol in figure 7.25. Just as the differential amplifier, it has two inputs, and the goal is to amplify the difference $v_i = v_p - v_n$, so that $v_o = A_v v_i$. Often, we define v_i in the other sense ($v_i = v_n - v_p$) and assume A_v is negative. Ideally, the amplifier has these characteristics:

1. The input impedance is infinite: $Z_i = \infty$. This means that currents i^- and i^+ at the inputs are always zero.
2. The output impedance is zero: $Z_o = 0$.
3. The voltage gain is infinite: $A_v = \infty$.[2]

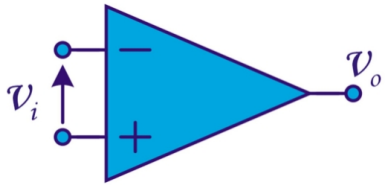


Figure 7.25

An OPAMP is almost never used in isolation, because with such a high gain, the output will almost certainly be at the supply voltage. Most often an OPAMP is used with other elements that apply feedback from output to the input. The concept of feedback will be studied in more detail in chapter 10.

[2] In practice, the gain will really be very high: ~ 100000.

Consider the topology in figure 7.26. The output of the OPAMP is fed back to the negative input terminal through resistance R_2 - this is called *negative feedback*. The input v_i is connected through resistance R_1 to the negative terminal of the OPAMP.

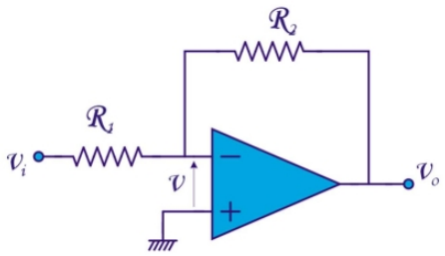


Figure 7.26

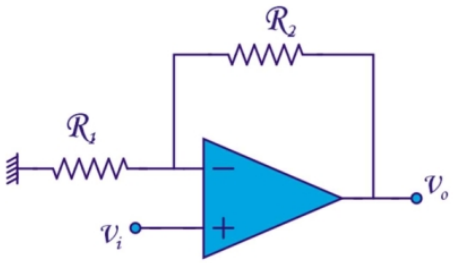


Figure 7.27

Voltage v is found by considering that all current through R_1 also goes through R_2 because the input impedance of the OPAMP is infinite:

$$\begin{aligned} \frac{v_i - v}{R_1} &= \frac{v - v_o}{R_2} \\ \Rightarrow (R_1 + R_2)v &= R_2 v_i + R_1 v_o \\ \Rightarrow v &= \frac{R_2 v_i + R_1 v_o}{R_1 + R_2} \end{aligned} \tag{7.18}$$

Output v_o is equal to $A_v v$, so:

$$\begin{aligned} v_o &= -A_v v \\ &= -A_v \frac{R_2 v_i + R_1 v_o}{R_1 + R_2} \\ \Rightarrow (R_1 + R_2)v_o &= -A_v R_1 v_o - A_v R_2 v_i \\ \Rightarrow v_o &= \frac{-A_v R_2}{R_1 + R_2 + A_v R_1} v_i \\ &\approx -\frac{R_2}{R_1} v_i \text{ if } A_v \to \infty \end{aligned} \tag{7.19}$$

Thus if A_v is very high, the gain of this topology does not depend on the OPAMP gain and is set by the ratio of resistors R_1 and R_2: $A = -\frac{R_1}{R_2}$. Because the gain is negative, this topology is called the *inverting amplifier*. The ratio $\frac{R_1}{R_2}$ can be set very precisely, and any temperature dependence disappears because both resistors vary in the same way with temperature. The gain $-\frac{R_2}{R_1}$ is the nominal gain A_n. Expression 7.19 can be rewritten as:

$$\begin{aligned} \frac{v_o}{v_i} &= -\frac{A_n}{1 + \frac{A_n}{A_v} + \frac{1}{A_v}} = \frac{A_v A_n}{A_n + A_v + 1} \\ &\approx -\frac{A_n}{1 + \frac{A_n}{A_v}} \end{aligned} \tag{7.20}$$

This expression is valid for all feedback topologies.

We can also compute $\frac{v}{v_i}$:

$$\begin{aligned}\frac{v}{v_i} &= \frac{v}{v_o}\frac{v_o}{v_i} \\ &= \frac{1}{A_v}\frac{-A_v R_2}{R_1 + R_2 + A_v R_1} \\ &\approx 0 \text{ if } A_v \to \infty\end{aligned} \tag{7.21}$$

If $A_v \to \infty$, the voltage $v \to 0$. This does not mean we can short-circuit both input terminals, but this realization often simplifies analysis. This is why the negative input is called a *virtual ground*.
For example, consider the circuit in 7.27, which is the same circuit as in 7.26, but with v_i applied at the positive terminal. If $v \to 0$, the voltage at the negative input node is also v_i and the current through R_1 is then equal to $\frac{v_i}{R_1}$. The same current flows through R_2:

$$v_o - v_i = R_2 \frac{v_i}{R_1}$$

and consequently:

$$v_o = \big(1 + \frac{R_2}{R_1}\big) v_i$$

In figure 7.26, we had an amplifier with a negative gain; here, we obtain an amplifier with a positive gain (the *non-inverting* amplifier).

7.4.2 The Real Amplifier

A real OPAMP, like the 741 OPAMP in figure 7.29 is far from ideal. The most important non-idealities are:

- The input impedance is not infinite, hence $i^-, i^+ \neq 0$. For example, the base currents when bipolar transistors are the inputs of the differential amplifier as first stage.
- There is an offset current $i_d = i^+ - i^-$, when $v_o = 0$. (typically $\sim 20nA$).
- Similarly, there is an offset voltage e_d that's required to make $v_o = 0$. (typically $\sim mV$) as in figure 7.28. This also means that when $v_d \approx 0$, then $v_o \approx E_{supply}$.
- The voltage gain is not infinite: $A_v \neq \infty$. (typically around 100 dB or 100000)
- The bandwidth ω_0 is limited.
- The slew rate and settling time. An OPAMP behaves as a two-pole system, so the output can not immediately follow the input. If we apply a step at the input, the output will increase but not immediately and will take some time to stabilize around the final value. These two effects are characterized by the slew rate and settling time, respectively.

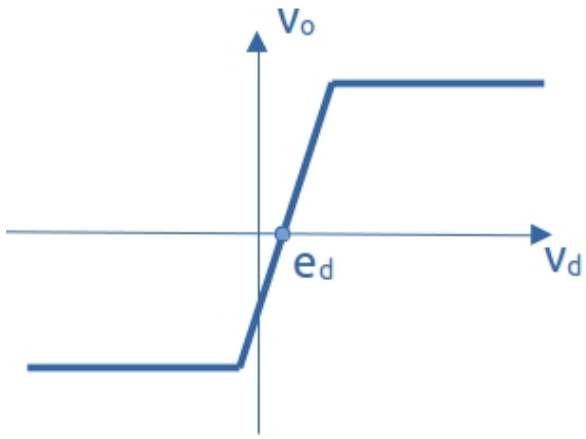


Figure 7.28

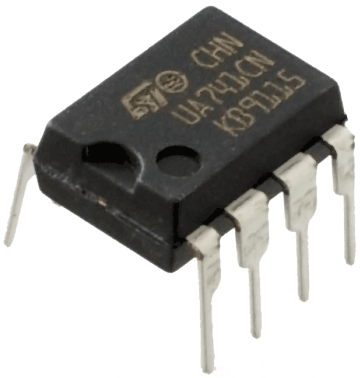


Figure 7.29

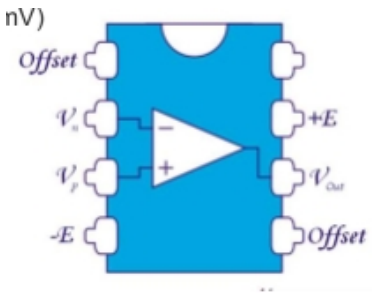


Figure 7.30

Furthermore, a real OPAMP requires a supply voltage $(+E, -E)$ and has a maximum allowed power consumption. The supply voltage can be symmetrical $(\pm 1.5V, \pm 5V, \pm 15V)$ or with the ground as reference $(+3.3V, +5V, +15V)$.
Figure 7.30 shows the typical pin configuration of a LM741 OPAMP, originally developed by Texas Instruments. Note the in -and output voltage v_p, v_n and v_o, the supply voltages $+E$ and $-E$ and two offset pins to adjust offset currents and voltage.

7.4.3 OPAMP Theory

To model these non-ideal effects, we will use the OPAMP model in figure 7.31, including a non-zero output impedance Z_o. To demonstrate this, we study the effect of input bias currents and offset voltages when the OPAMP is used in the simple circuit of figure 7.26. Replacing the OPAMP by its model (with $Z_o = 0$) gives the equivalent circuit in figure 7.32. The idea to solve this problem is to consider all sources individually (with all other sources = 0), and adding the results at the end. This is an application of the superposition principle. When we suppose $A_v \to \infty$:

- $v_i : v_0 = -\frac{R2}{R1}\, v_i$,
- $e_d : v_o = \left(1 + \frac{R2}{R1}\right) e_d$, because this situation is as in figure 7.27 with voltage e_d at the

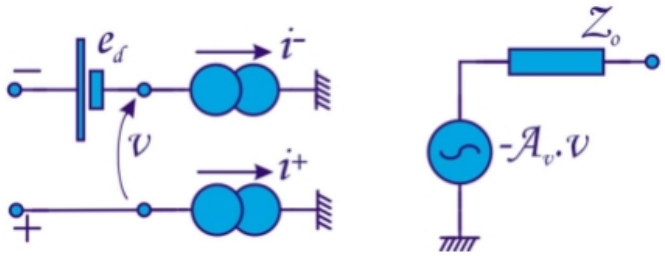


Figure 7.31

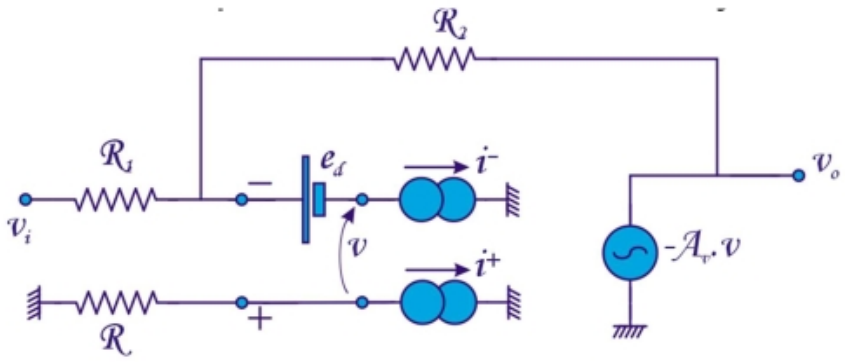


Figure 7.32

negative input instead of v_i.

- i^- : $v_o = R_2\, i^-$ because with Millman:

$$v = \frac{G_2 v_o - i^-}{G_1 + G_2}$$

and since $v_o = -A_v v$, we rearrange and obtain:

$$v_o = \frac{A_v}{G_1 + G_2 + A_v G2} i^- \approx R_2\, i^-$$

Another way to see this, is to realize there is no current through R_1: because $v = 0$, both terminals of R_1 are at zero voltage. All current drawn by i^- goes through R_2, thus $v_o - v = R_2\, i^-$.

- i^+ : $v_o = -R\left(1 + \frac{R2}{R1}\right)\, i^+$ because a current i^+ generates a voltage $-R\, i^+$ at the positive terminal, replicating the situation of figure 7.26;

The final expression is thus:

$$v_o = -\frac{R_2}{R_1} v_i + \left(1 + \frac{R_2}{R_1}\right)\, e_d + R_2\, i^- - R\left(1 + \frac{R_2}{R_1}\right)\, i^+$$

We can draw these conclusions:

- If e_d is of the same order of magnitude as v_i, it will have a large impact on the output. The best way to take e_d into account, is to first measure the output voltage without applying v_i. In this way, you measure $\left(1+\frac{R2}{R1}\right) e_d$ and you subtract this value when you measure v_o with v_i at the input.
- The impact of i^- can be reduced by choosing a small R_2. This is feasible because A_n is the ratio of resistors; their actual value is of less importance.
- If you choose $R = R_1||R_2$, then $R_2\ i^- - R\left(1+\frac{R2}{R1}\right)\ i^+ = R_2(i^- - i^+)$ and you significantly reduce the impact of the input bias currents.

Since the gain is not infinite, a relative gain error ϵ will always be present. If we call the true gain of the OPAMP with feedback circuitry A, we know from equation 7.20 that:

$$A = \frac{v_o}{v_i} \approx -\frac{A_n}{1+\frac{A_n}{A_v}} \approx -A_n \tag{7.22}$$

The relative gain error ϵ is thus:

$$\begin{aligned} \epsilon &= \frac{A_n - A}{A_n} = \frac{A_n - \frac{A_n}{1+\frac{A_n}{A_v}}}{A_n} \\ &= 1 - \frac{1}{1+\frac{A_n}{A_v}} = 1 - \frac{A_v}{A_v + A_n} \\ &= \frac{A_n}{A_v + A_n} \approx \frac{A_n}{A_v} \end{aligned} \tag{7.23}$$

The gain error thus increases with increasing A_n and with decreasing OPAMP gain (which happens at higher frequencies).

The effect of feedback on the bandwidth

In reality, the OPAMP has a limited bandwidth. We will model the OPAMP as a first-order system with a pole in $\omega = \omega_0$:

$$A_v = \frac{A_{v0}}{1 + j\frac{\omega}{\omega_0}}$$

This system has a cut-off frequency $f_0 = \omega_0/2\pi$.
Substituting this expression in equation 7.22, gives:

$$\begin{aligned} A &\approx -\frac{A_n}{1+\frac{A_n}{A_v}} && \text{equation 7.22} \\ &= -\frac{A_n}{1+\frac{A_n}{A_{v0}}\left(1+j\frac{\omega}{\omega_0}\right)} && \text{substitution} \\ &= -\frac{A_n}{\left(1+\frac{A_n}{A_{v0}}\right)+j\frac{\omega}{\omega_0}\left(\frac{A_n}{A_{v0}}\right)} && \text{rearrange real - imaginary parts} \\ &\approx -\frac{A_n}{\left(1+\frac{A_n}{A_{v0}}\right)}\frac{1}{1+j\frac{\omega}{\omega_0}\left(\frac{A_n}{A_{v0}}\right)} && \text{because } \left(\frac{A_n}{A_{v0}}\right)^2 \text{ is very small} \\ &\approx -\frac{A_n}{1+j\frac{\omega}{\omega_n}} \end{aligned}$$

where ω_n is a new cut-off frequency:

$$\omega_n = \frac{A_{v0}\,\omega_0}{A_n}$$

This means that by using the OPAMP with the feedback circuitry, the gain has decreased and becomes A_n, but the bandwidth (the cut-off frequency) has increased. What remains constant is the *gain-bandwidth product* GBW:

$$A_n \times \omega_n = A_{v0} \times \omega_0 = GBW$$

Figure 7.33 represents the Bode curve of the OPAMP self (in blue), with a cut-off at ω_0 and DC-gain A_{v0}, and the OPAMP with feedback circuitry (in green). The latter curve has a lower DC-gain A_n, but a larger bandwidth ω_n. Another issue is also highlighted in the graph: if the frequency increases, the gain decreases and the relative gain error increases. If a maximum ϵ is imposed, the signal frequency cannot go beyond ω_{max}.

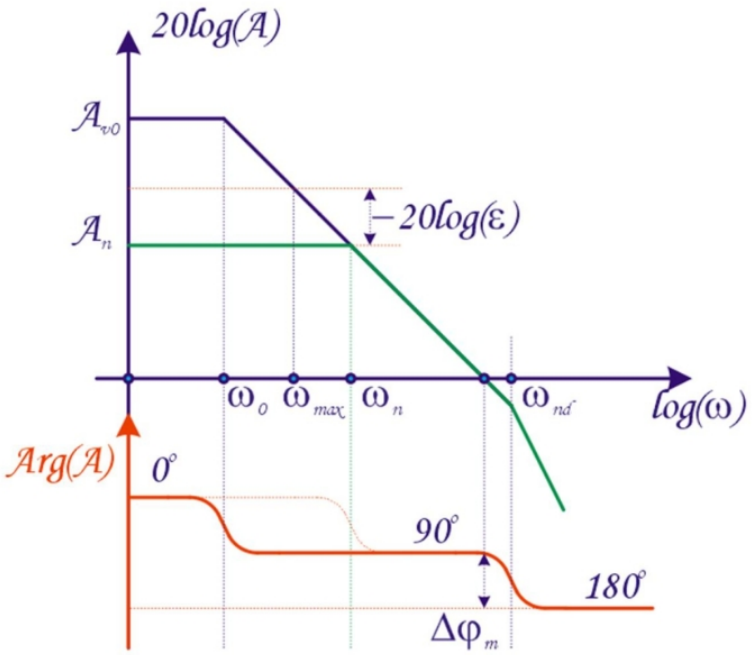


Figure 7.33

Summary

When an OPAMP is used for data acquisition (which is often the case), lots of errors will be introduced:

- due to an offset voltage e_d,
- due to a differential input current i_d,
- due to the limited gain $A_v < \infty$, which results in a relative gain error ϵ,

- due to the limited bandwidth ω_0

You'll have to take all these errors into account, or avoid them by e.g. using OPAMPs with a low offset. In any case, because an OPAMP is inherently a second-order system, we have to live with a delay during acquisition, due to a non-zero settling time.

7.4.4 Unity Gain Buffer

A commonly used OPAMP configuration is the one in figure 7.34, where the output node is directly connected to the negative input terminal. With v equal to zero, we immediately see that $v_o = v_i$, or that $A_n = 1$. We use the equivalent model in figure 7.35 to study the different parameters more formally.

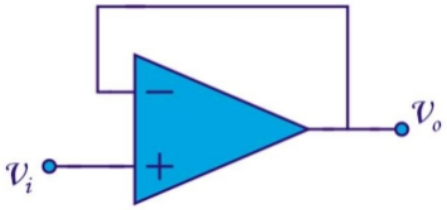


Figure 7.34

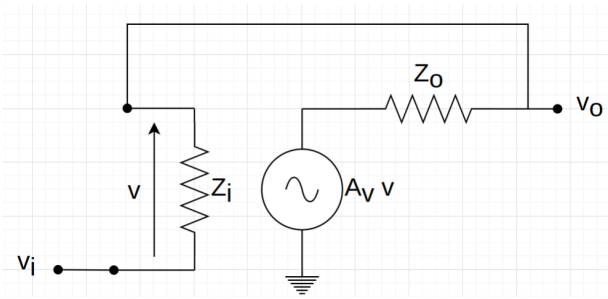


Figure 7.35

1. Voltage gain A_v:

$$\begin{aligned} v_o &= \frac{G_o(-A_v v) + G_i v_i}{G_o + G_i} \\ &= \frac{G_o A_v(v_i - v_o) + G_i v_i}{G_o + G_i} \\ \rightarrow \frac{v_o}{v_i} &= \frac{G_i + A_v G_o}{G_o + G_i + A_v G_o} \\ &\approx \frac{A_v G_o}{A_v G_o} = 1 \end{aligned}$$

2. Input impedance Z_i:

$$\begin{aligned} i_i &= G_i(v_i - v_o) \\ &= G_i(1 - \frac{G_i + A_v G_o}{G_o + G_i + A_v G_o})v_i \\ &= G_i \frac{G_o}{G_o + G_i + A_v G_o} v_i \\ &\approx \frac{G_i}{A_v} v_i \\ \rightarrow Z_i = \frac{v_i}{i_i} &= A_v \; Z_i \end{aligned}$$

3. Output impedance Z_o:

$$\begin{aligned} i_o &= G_i v_o + G_o(v_o - (-A_v v)) \\ &= (G_i + (1 + A_v)G_o)v_o \\ &\approx A_v G_o v_o \\ \rightarrow Z_o = \frac{v_o}{i_o} &= \frac{Z_o}{A_v} \end{aligned}$$

So this circuit has unity gain, a very high input impedance $A_v Z_i$ and a very low output impedance $\frac{Z_o}{A_v}$. In summary, it is the ideal buffer to isolate one stage from the next.

Chapter 8

Transistors at High Frequency

In the BJT small-signal model of figure 6.35, there are two capacitors, C_π and C_μ. These capacitors model the pn-junctions between base and emitter and base and collector, i.e. they are caused by the depletion regions. Because the emitter-base junction is forward biased and the base-collector junction is reversed biased, the depletion zone in the former junction is a lot larger than in the latter, and thus $C_\pi \gg C_\mu$.
We first analyze the impact of C_π on the behavior of the BJT transistor at high frequencies. Next, we study the impact of C_μ. To do this, we replace C_μ by a Miller capacitor C_M to simplify the analysis.

8.1 Giacoletto Model at High Frequencies

In this section, we assume that $C_\pi \gg C_\mu$ and that $C_\mu \approx 0$. The small-signal model under these assumptions is shown in figure 8.1. With an input current i_b, the base-emitter voltage is:

$$v_{be} = \frac{r_\pi}{1 + j\omega r_\pi C_\pi} i_b$$

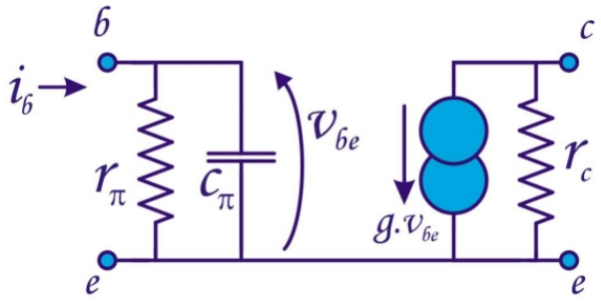


Figure 8.1

with $Z_\pi = \frac{r_\pi}{1+j\omega r_\pi C_\pi}$ the parallel combination of r_π and C_π. This impedance is equal to r_π for low frequencies, but begins to decrease at a frequency f_β:

$$2\pi f_\beta = \omega_\beta = \frac{1}{r_\pi C_\pi}$$

Beyond this frequency the base starts to degrade and the transistor capacity to amplify the input current decreases.
However, even beyond f_β the transistor can still be used. He stops working when there is no current amplification, i.e. when $\frac{i_c}{i_b} = 1$, where i_c is the current between collector and emitter when we short-circuit them (in the AC equivalent circuit, obviously). This short-circuit current gain $A_{i,sc}$ is the ratio of the current generated by the dependent current source with respect to the input current:

$$A_{i,sc} = \frac{i_c}{i_b} = g\, v_{be} = g\frac{r_\pi}{1 + j\omega r_\pi C_\pi}$$

We compute the pulsation ω_T when $|A_{i,sc}| = 1$:

$$\left| g\frac{r_\pi}{1 + j\omega r_\pi C_\pi} \right| = 1$$
$$= \frac{g r_\pi}{\sqrt{1 + \frac{\omega^2}{\omega_\beta^2}}} = \frac{\beta}{\sqrt{1 + \frac{\omega^2}{\omega_\beta^2}}}$$

because $g\, r_\pi = \frac{I_{CQ}}{v_{th}} \frac{v_{th}}{I_{BQ}} = \frac{I_{CQ}}{I_{BQ}} = \beta$. Consequently:

$$\omega_\beta^2 \beta^2 = \omega_\beta^2 + \omega^2$$
$$\Rightarrow \omega = \omega_\beta \sqrt{\beta^2 - 1} \approx \omega_\beta \beta$$

This pulsation $\omega_T = \beta\ \omega_\beta$ is the pulsation beyond which the BJT no longer amplifies the current and thus becomes useless.

8.2 The Miller Capacitor

In this section, we will also take C_μ into account. We study the common-emitter amplifier from figure 7.8 when the input signal v_i has high-frequency components. We assume that both C_B and C_E are short circuits, i.e. we are in the correct working domain (the right part of the Bode curve in figure 7.7). With both C_π and C_μ present, and an output impedance R_S of the signal source, we obtain the AC circuit from figure 8.2.

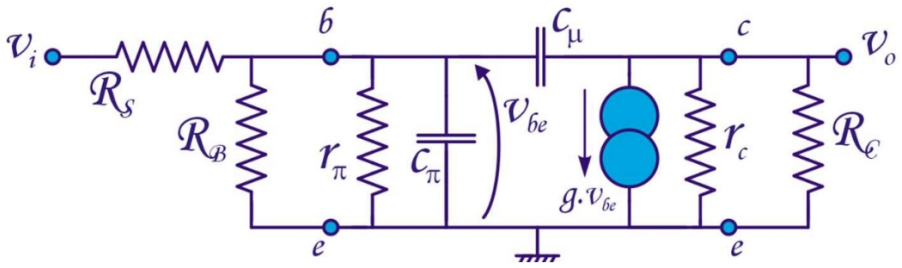


Figure 8.2

The difficulty in this circuit is the presence of C_μ: this capacitor couples the base to the collector, making the analysis hard. To proceed, we want to remove C_μ and replace it by

something else, like an element in parallel with C_π. This is where the Miller capacitor comes in.
The key insight is that we can replace the series capacitor C in figure 8.3, which induces a current $i = j\omega C(v_2 - v_1)$, by the two loops in figure 8.4. From an electrical point of view, these circuits are identical because the currents and voltages at the in- and output terminals are identical.

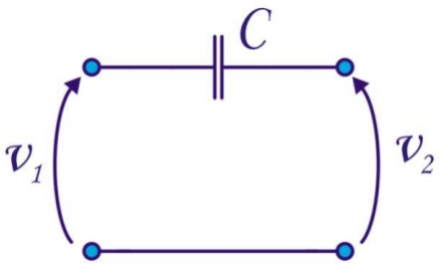


Figure 8.3

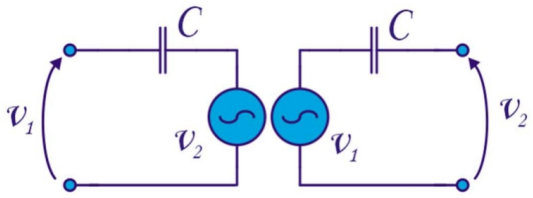


Figure 8.4

In the right loop, we replace the voltage source with its Norton equivalent $i_N = j\omega C v_1$, as in figure 8.5.

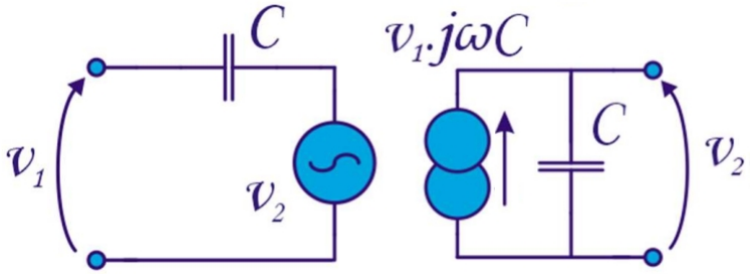


Figure 8.5

If we replace C_μ in this way in the AC-circuit of figure 8.2, we obtain the circuit in figure 8.6. Note that Z_π is $r_\pi||C_\pi||R_B$ and $R_{eq} = r_c||R_C$. In this circuit, we will now:

- Neglect current source $v_{be}j\omega C_\mu$ when it is dominated by $g\ v_{be}$. This is valid when:

$$\begin{aligned} v_{be}j\omega C_\mu &\ll g\ v_{be} \\ |j\omega C_\mu| &\ll |g| \\ \rightarrow \omega &\ll \omega_1 = \frac{g}{C_\mu} \end{aligned}$$

  This is the first criterion: $\omega \ll \frac{g}{C_\mu}$

- Neglect C_μ in the output loop because its impedance is lot larger then R_{eq}. This is valid if:

$$\begin{aligned} \frac{1}{j\omega C_\mu} &\gg R_{eq} = r_c||R_C \\ \rightarrow \omega &\ll \omega_2 = \frac{1}{R_{eq}C_\mu} \end{aligned}$$

  This is the second criterion: $\omega \ll \frac{1}{R_{eq}C_\mu}$

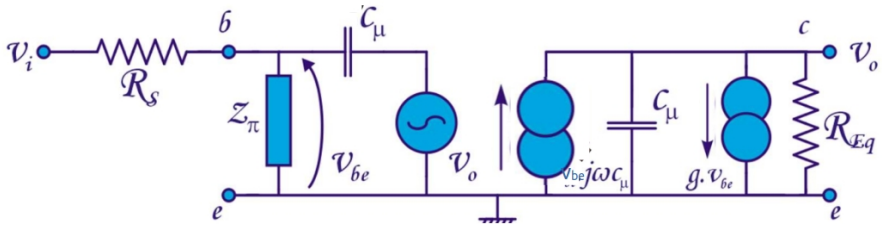


Figure 8.6

Note that in a typical scenario, e.g. with $I_{CQ} = 1$ mA and $R_C = 1$ kΩ, $g \gg G_C$ and $g \gg g_c$ (this is always valid) and thus mostly $g \gg G_c + g_c = \frac{1}{R_{eq}}$. So if the second criterion is valid, the first will be valid as well. In general, the Miller conditions are satisfied.
After neglecting $j\omega\ C_\mu\ v_{be}$ and C_μ, the small-signal circuit becomes the one in figure 8.7. In this circuit, we know that $v_o = -g\ R_{eq}\ v_{be}$. Hence, the current though C_μ in the left loop is equal to:

$$\begin{aligned} i_{C_\mu} &= j\omega C_\mu (v_{be} - v_o) \\ &= j\omega C_\mu (1 + g\ R_{eq})\ v_{be} \end{aligned}$$

So we can replace C_μ and the source that provides v_o by a single capacitor with capacitance $C_\mu(1 + g\ R_{eq})$. This is the Miller capacitor:

$$C_M = C_\mu (1 + g R_{eq}) \tag{8.1}$$

as in figure 8.8.

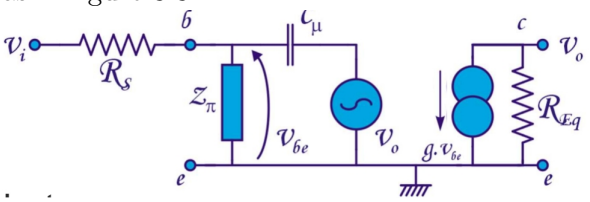


Figure 8.7

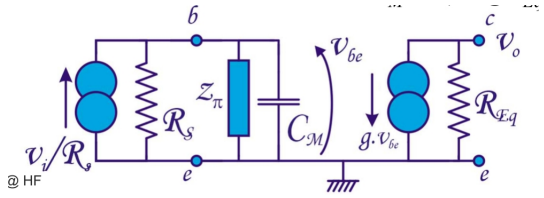


Figure 8.8

We can in the left loop of figure 8.8 group al impedance in a single $Z_{eq} = \frac{R}{1+j\omega R(C_\pi + C_M)}$ with $R = r_\pi || R_B || R_S$ the parallel combination of all resistors in the input part of the circuit, and $C_\pi + C_M$ the parallel combination of C_π and C_M. The base-emitter voltage is then equal to:

$$v_{be} = \frac{1}{R_S} \frac{R}{1 + j\omega R(C_\pi + C_M)} v_i$$

and the voltage gain is:

$$A_v = \frac{v_o}{v_i} = -gR_{eq}\frac{v_{be}}{v_i} = -gR_{eq}\frac{1}{R_S}\frac{R}{1 + j\omega R(C_\pi + C_M)}$$

When R_S is small, $R \approx R_S$ and:

$$A_v \approx -g \frac{R_{eq}}{1 + j\omega R(C_\pi + C_M)} \tag{8.2}$$

8.3 Miller's Theorem

Equation 8.1 is an instance of Miller's theorem, which aims to replace a floating impedance with two grounded impedances as we did in figure 8.3. For the derivation, refer to figure 8.9.

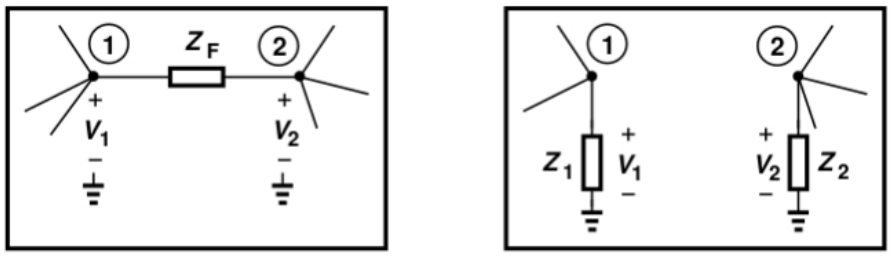


Figure 8.9

We wish to transform Z_F to two grounded impedances as depicted in figure 8.9(b), while ensuring all of the currents and voltages in the circuit remain unchanged. This means:

$$\frac{V_2 - V_1}{Z_F} = \frac{V_1}{Z_1}$$
$$\frac{V_1 - V_2}{Z_F} = -\frac{V_2}{Z_2}$$

If we denote the voltage gain from node 1 to node 2 as $A_v = \frac{V_2}{V_1}$, we find:

$$\begin{aligned} Z_1 &= \frac{V_1}{V_2 - V_1} Z_F \\ &= \frac{1}{1 - A_v} Z_F \end{aligned}$$

for Z_1 and

$$\begin{aligned} Z_2 &= \frac{V_2}{V_1 - V_2} Z_F \\ &= \frac{1}{1 - A_v^{-1}} Z_F \end{aligned}$$

for Z_2.

8.4 Conclusion

At high frequencies, we must take the parasitic capacitances of the transistor into account. For floating capacitances like C_μ, we use Miller's theorem and replace it by C_M. Note that we can only do this if the Miller conditions are satisfied:

1. $\omega \ll \frac{g}{C_\mu}$
2. $\omega \ll \frac{1}{R_{eq} C_\mu}$

From equation 8.2, we see that there is a cut-off frequency for the gain:

$$\omega_H = \frac{1}{c_\pi + C_M}$$

Above this frequency, the gain A_v of the common-emitter amplifier begins to decrease, and we can extend figure 7.7 to include this cut-off frequency, as in figure 8.10. We also know that $\omega_H < \frac{1}{r_\pi C_\pi} = \omega_\beta$. This means that performance degradation is initially due to a direct signal path from base to collector through C_μ, and that base degradation happens later (i.e. at higher frequencies).

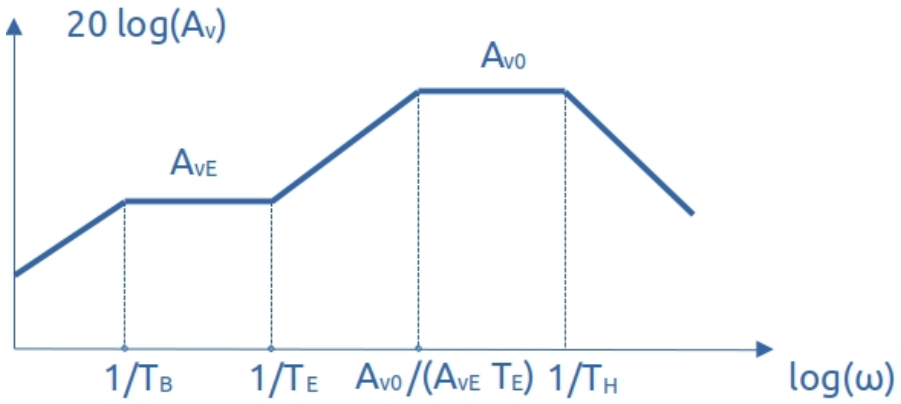


Figure 8.10

Chapter 9

Power Amplifiers

In the previous chapter, we studied all kinds of amplifiers. However, we never looked at power amplification and how power was provided to the load resistance, but we only were interested in amplification of the input voltage. In this chapter, we're mostly interested in power consumption and efficiency, and we'll look at amplifiers of class A, B, C, D and S. We'll also study current amplifiers like push-pull amplifiers, and selective amplifiers that only amplify signals around a central frequency.

9.1 Introduction

Until now, we were not concerned with power consumption or efficiency of the amplifiers we have studied. We only cared about achieving high gain. In this chapter, we look at *power amplifiers*, amplifiers made to deliver power - and to do this as efficiently as possible.
First, we will consider a simple common emitter amplifier in figure 9.1. Notice that we don't consider an emitter resistance R_E. This is because typically, R_E is quite low; if not, the possible voltage swing along the dynamic load line would become too small.
With proper biasing and neglecting $V_{CE,Sat}$, we see that $V_{CEQ} = \frac{E}{2}$ and $I_{CQ} = \frac{E}{2R_C}$ so that Q is nicely in the middle of the operating domain, as in figure 9.2. The load line is given by $E = R_C\, i_c + v_{CE}$. Furthermore, we assume that the collector resistor R_C is the load to which we want to deliver power.

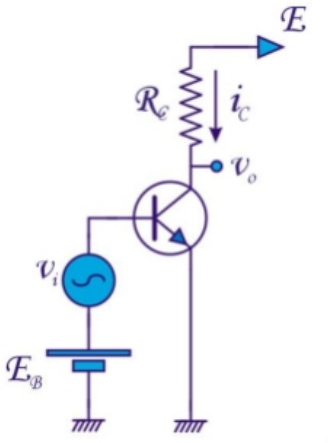


Figure 9.1

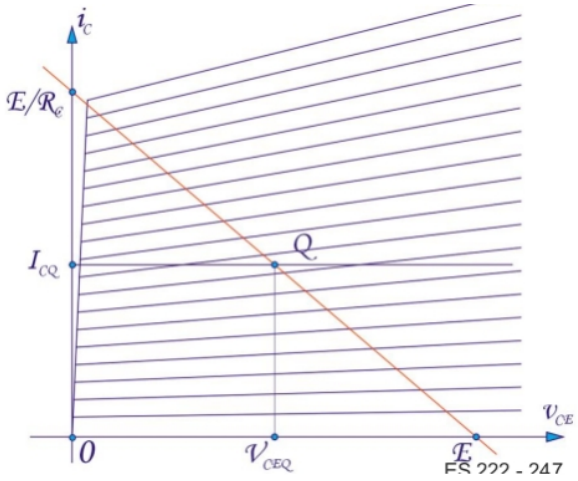


Figure 9.2

We assume that the applied input signal v_i is sinusoidal. If E_B is large enough, the collector current i_C will also be a sinusoid and the transistor is always conducting, as in the top of figure 9.3. But if we decrease the input bias source E_B, I_{CQ} can be reduced, and from a certain point on (namely when $E_B + v_i = V_{BEQ}$), i_C will become equal to zero (figure 9.4 bottom), and can't become negative because the circuit can't conduct in the other direction. If E_B decreases even further, the transistor will conduct less then half of the time (figure 9.4). The *conduction angle* θ is defined as:

$$\theta = \frac{T_{conduct}}{T} \times 180^\circ \tag{9.1}$$

Based on the conduction angle, 4 different classes of amplifiers are defined:

1. Class A: $\theta = 180^\circ$
2. Class AB: $90^\circ < \theta < 180^\circ$
3. Class B: $\theta = 90^\circ$
4. Class C: $\theta < 90^\circ$

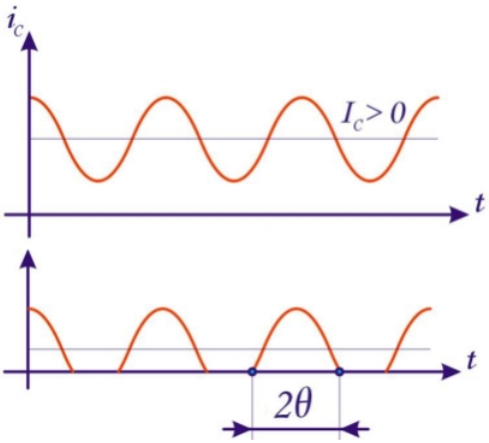


Figure 9.3

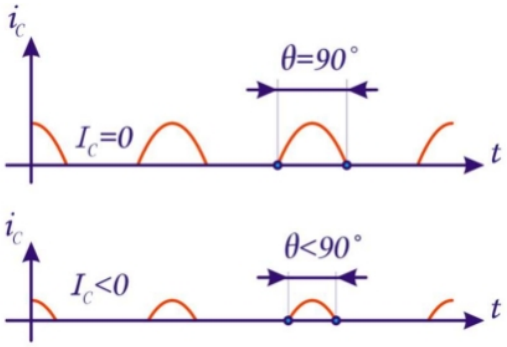


Figure 9.4

In the rest of this chapter, we study a circuit for each type of conduction. We will compute the power efficiencies, by considering only the circuit on the collector side. All other power consumptions, like base currents or power consumed in the emitter (source) resistance will be neglected.

The instantaneous power consumed by an element in a circuit is $p(t) = v(t)i(i)$ with v and i the instantaneous voltage and current through the element. When the signals are sinusoidal, we define the average power dissipated during one period T:

$$P = \frac{1}{T} \int_{t_0}^{t_0+T} v(t)\, i(t)\, dt$$

9.2 Class A Amplifier

We assume that the applied input signal v_i is sinusoidal and thus:

$$v_{CE} = V_{CEQ} + V_{cem} \sin(\omega t) = V_{CEQ} + v_{ce}(t)$$
$$i_C = I_{CQ} - I_{cm} \sin(\omega t) = I_{CQ} + i_c(t)$$

with $V_{cem} = R_C I_{cm}$.
Because Q is located in the middle of the operating region, we also have $V_{CEQ} = E/2$ and $I_{CQ} = \frac{E}{2R_C}$. Consequently, the voltage and current amplitudes are limited to: $V_{cem} \leq E/2$ and $I_{cem} \leq \frac{E}{2R_C}$.
We calculate[1]:

- P_D, the power delivered to the circuit by the supply E:

$$P_D = \frac{1}{T} \int_{t_0}^{t_0+T} E\ i_C(t)\ dt = E\ I_{CQ} = \frac{E^2}{2R_C}$$

- P_C, the power dissipated by the transistor:

$$\begin{aligned} P_C &= \frac{1}{T} \int_{t_0}^{t_0+T} v_{CE}(t)\ i_C(t)\ dt \\ &= \frac{1}{T} \int_{t_0}^{t_0+T} (V_{CEQ} + V_{cem}\sin(\omega t))\ (I_{CQ} - I_{cm}\sin(\omega t))\ dt \\ &= \frac{1}{T} \int_{t_0}^{t_0+T} (V_{CEQ}\ I_{CQ} - V_{cem}\ I_{cm}\sin(\omega t)^2)\ dt \\ &= V_{CEQ}\ I_{CQ} - \frac{1}{2} V_{cem}\ I_{cm} = \frac{E^2}{4R_C} - \frac{1}{2}\frac{V_{cem}^2}{R_C} \\ &= \frac{E^2}{4R_C} - \frac{1}{2}\frac{E^2}{4R_C} = \frac{E^2}{8R_C} \text{ at } V_{cem} = E/2 \text{ and } I_{cm} = \frac{E}{2R_C} \end{aligned}$$

- P_L, the AC power delivered to the load R_C because the only useful power delivered to the load is the variation of the signal around the average value. Note that $v_{R_C} + v_{ce} = 0$ and $i_{R_C} = i_c$. So:

$$\begin{aligned} P_L &= \frac{1}{T} \int_{t_0}^{t_0+T} v_{R_C}(t)\ i_{R_C}(t)\ dt = \frac{1}{T} \int_{t_0}^{t_0+T} v_{ce}(t)\ i_c(t)\ dt \\ &= \frac{1}{2} V_{cem}\ I_{cm} \\ &= \frac{1}{2}\frac{E^2}{4R_C} = \frac{E^2}{8R_C} \text{ at } V_{cem} = E/2 \text{ and } I_{cm} = \frac{E}{2R_C} \end{aligned}$$

- P_J, the rest:

$$\begin{aligned} P_J = P_D - P_C - P_L &= \frac{E^2}{2R_C} - \left(\frac{E^2}{4R_C} - \frac{1}{2}\frac{V_{cem}^2}{R_C}\right) - \frac{1}{2} V_{cem}\ I_{cm} \\ &= \frac{E^2}{4R_C} = R_C I_{CQ}^2 \end{aligned}$$

This last term P_J is caused by the DC current I_{CQ} in the load resistor R_C.
The efficiency of the amplifier η is defined as the ratio between power delivered to the load over the total power delivered by the supply:

$$\eta = \frac{P_L}{P_D} = \frac{V_{cem}^2}{E^2} \tag{9.2}$$

[1]Note that $\int_0^T \sin(\omega t)\ dt = 0$ and $\int_0^T \sin^2(\omega t)\ dt = \frac{T}{2}$

With $V_{cem,max} = \frac{E}{2}$ we find the maximum efficiency: $\eta_{max} = \frac{1}{4}$. The quality factor F is the ratio between the maximum power consumed by the transistor $P_{C,max}$ and the maximum power consumed by the load $P_{L,max}$. The former happens when $V_{cem} = 0$ and is equal to: $P_{C,max} = \frac{E^2}{4R_C}$, the latter occurs at full swing $V_{cem} = E/2$ and is equal to $P_{L,max} = \frac{E^2}{8R_C}$. Thus

$$F = \frac{P_{C,max}}{P_{L,max}} = 2$$

These different power contributions are shown in figure 9.5 as a function of the voltage swing

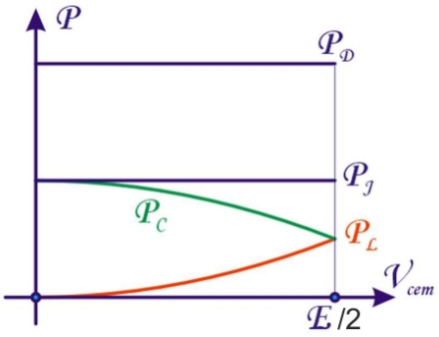


Figure 9.5: Power distribution for increasing V_{cem}

V_{cem}. Note how the maximum efficiency is only 25%, achieved at the maximal value of V_{cem}. Half the power $P_J = \frac{E^2}{8R_C} = R_C I_{CQ}^2$ is used to generate the bias current for the transistor, which also runs trough the load resistor R_C but doesn't provide anything useful. This power is useless and should be eliminated to increase the efficiency η.
To improve this, we want to avoid any DC current in the load. We propose the circuit in figure 9.6. If L is very large, the inductor will be a short-circuit for DC signals. And if C_L is very large, the load resistor R_L is isolated from the bias currents. Thus I_{CQ} doesn't flow through the load, in contrast with the previous situation. Consequently, $V_{CEQ} = E$ and $I_{CQ} = \frac{E}{R_C}$ and the operating point Q is uniquely determined.

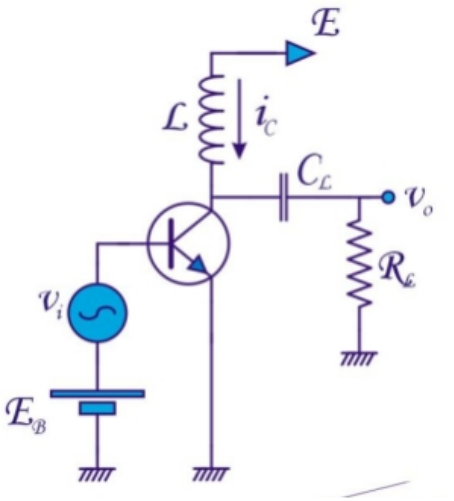


Figure 9.6: Class A Amplifier

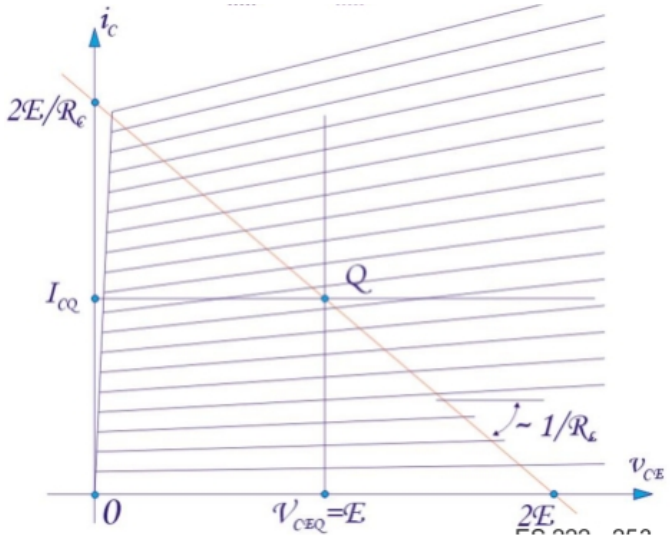


Figure 9.7: Load lines and Q-point

For AC signals however, the inductor is an open circuit and C_L is a short circuit, so we can write the dynamic load line: $v_{ce} = -R_L\ i_c$. The load lines (static in blue, dynamic in red) are sketched in figure 9.7. Note how v_{CE} can become higher then the power supply E. This is because the inductance will create an electromotive force (EMF) at the frequency we are working at to keep the current through it constant. This EMF can increase the voltage at the collector to $2E$. This also means that $V_{cem,max} = E$ and $I_{cem,max} = E/R_L$.
We can recompute the different powers:

- $P_D = \frac{1}{T}\int_{t_0}^{t_0+T} E\ i_C(t)dt = E\ I_{CQ} = \frac{E^2}{R_L}$
- $P_C = \frac{1}{T}\int_{t_0}^{t_0+T} v_{CE}(t)\ i_C(t)\ dt = V_{CEQ}\ I_{CQ} - \frac{V_{cem}^2}{2R_L} = \frac{E^2}{2R_L}$
- $P_L = \frac{1}{T}\int_{t_0}^{t_0+T} v_{ce}(t)\ i_c(t)\ dt = \frac{V_{cem}^2}{2R_L} = \frac{E^2}{2R_L}$

Note how $P_C + P_L = P_D$ and thus $P_J = 0$. This is because there runs no DC (and hence useless) power through the load resistance. The different powers are shown in figure 9.8.

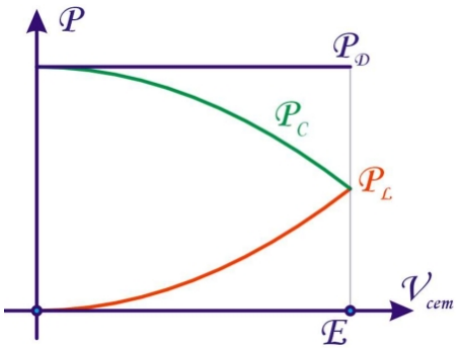


Figure 9.8: Power distribution for class A amplifier

The efficiency $\eta = \frac{P_L}{P_D} = \frac{V_{cem}^2}{2E^2}$ so that $\eta_{max} = \frac{1}{2}$, and the quality factor $F = 2$. The maximum efficiency (reached at maximum amplitude) is thus 50%, instead of 25% as before.

9.2.1 Improvement to the class A amplifier

This issue with the previous circuit is this: typically, the power supply E is given, and when you buy a speaker, the internal resistance R_L is also given. This means that the maximum power you can deliver to the load, namely $\frac{E^2}{2R_L}$ is also fixed, even though the speaker may have a higher $P_{L,max}$.
An alternative circuit to the one in figure 9.6 is the one in figure 9.9 where a transformer is used to transfer the AC power to the load and the conversion factor n can be set. This additional degree of freedom will allow to increase the maximum power.

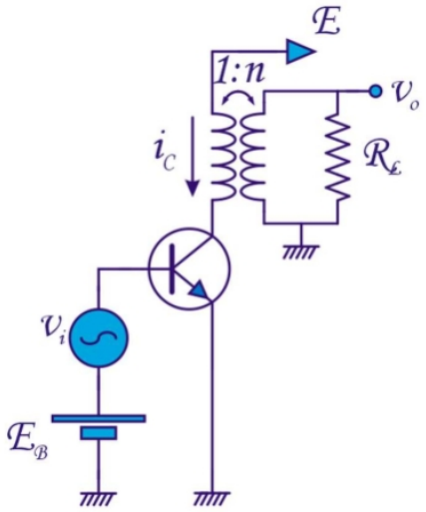


Figure 9.9: Improved Class A Amplifier

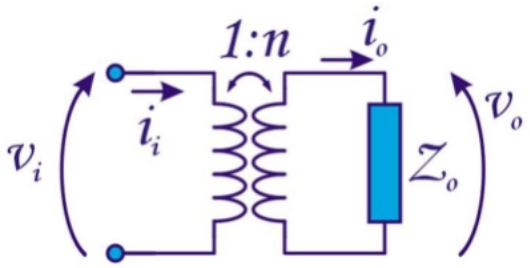


Figure 9.10: Transformer

In this circuit, we have the same DC load line as before: $v_{CE} = E$, but the AC load line is now $v_{ce} = -R'_L\, i_e$ where R'_L is the load as seen by the biasing circuit.
To understand this, consider the transformer in figure 9.10. A transformer works by generating a changing magnetic field $\Phi_B(t)$ with an inductor. This flux flows then through another inductor which generates an EMF due to Faraday's law of induction: $\mathcal{E} = -\frac{d\Phi_B}{dt}$. Because a transformer only works for AC signals, the load seen by the DC circuit is zero - the transformer acts as a short-circuit. For AC signals however, as in figure 9.10, the impedance Z_i seen by the circuit is different. Because the relation between currents and voltages is:

$$\begin{aligned} n\, I_o &= I_i \\ V_o &= n\, V_i \end{aligned}$$

we find that $Z_i = \frac{V_i}{I_i} = \frac{V_o}{n^2\, I_o} = \frac{Z_o}{n^2}$. This means that the load R_L is reflected into the circuit and the circuit sees $R'_L = R_L/n^2$. For a given E, R_L and $P_{L,max}$, we know for the reflected load (the apparent resistor) that $P_{L,max} = \frac{E^2}{2R'_L}$ and thus:

$$R'_L = \frac{E^2}{2P_{L,max}}$$

and from R'_L and the true R_L, we find the turns ratio n:

$$n = \sqrt{\frac{R_L}{R'_L}}$$

In an audio amplifier, this turns ratio can often be set be turning a screw.

9.3 Class B Amplifier

In the class A amplifier from the previous section, there was always a current I_{CQ} and a current variation i_c around this value because there is a bias source E_B that lifts the base voltage at the input transistor to a level that allows conduction in the right branch. This also

means that the transistor dissipates power even when there is no signal ($V_{cem} = 0$) - it will even dissipate all power delivered to the circuit, which is constant.
We could remove this bias source, as in figure 9.11. However, in that case there will only be a non-zero voltage of R_L during half of the period, namely when $v_i > 0$, as in figure 9.12. Note that the voltage is negative because the circuit has a negative gain.

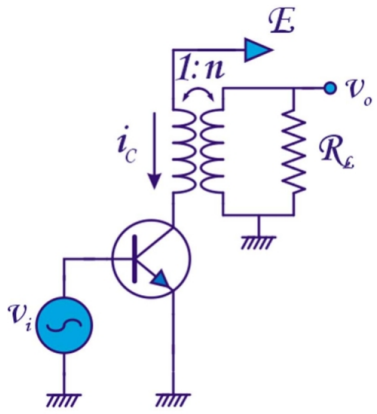


Figure 9.11: Class A Amplifier without E_B

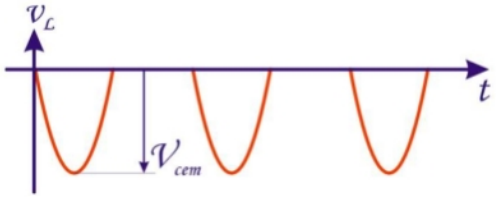


Figure 9.12: Signal produced in R_L

This is off course not an acceptable situation, but we can create a signal during the entire period of the input signal by effectively putting two of these circuits on top of each other, as in figure 9.13. Note that we use both v_i and $-v_i$ and two npn-transistors with interconnected emitters and opposite base-emitter voltages. On the load side, note that the dots signify wether the turns of the inductors turn in the same direction or not.
Transistor ① will conduct only when $v_i > 0$ and generates a signal of the form of figure 9.12 in R_L. Transistor ② conducts only when $v_i < 0$ and generates the opposite signal of figure 9.12. When both signals are added together, the complete wave form is restored. Note that we have neglected that $V_{BEQ} \approx 0.6$ V - we will come back to this later.
The DC load line for transistor ① has the same expression as behavior: $v_{CE} = E$, but $I_{CQ} = 0$ because the transistor is blocked when $v_i = 0$. Thus the operating point lies on the x-axis of figure 9.14 and the dynamic load line with slope $-\frac{1}{R'_L}$ passes through this point. As v_i increases, v_{CE} of transistor ① decreases because we move up the load line, as shown in figure 9.14, until the maximum current E/R'_L is reached. At the same time, transistor ② is blocked but v_{CE} of ② increases from E to $2E$ because the transformer induces an EMF from the right inductor to the inductor in the loop with the emitter-collector of transistor ②. The V_{CE} of ② moves along the horizontal line from Q to $2E$. When v_i becomes negative, the situation is reversed for both transistors.

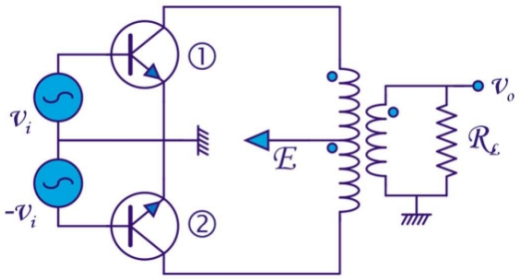


Figure 9.13: Class B Amplifier

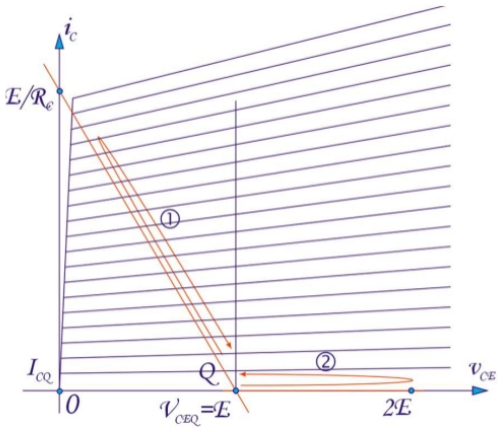


Figure 9.14: Load lines of figure 9.13

Let's compute the different powers, just as before. However, keep in mind that every transistor is only active half of the time, so we integrate during $T/2$ and multiply the result by two:

- $P_D = 2\frac{1}{T}\int_{t_0}^{t_0+T/2} E\ I_{cm}\ \sin(\omega t)dt = \frac{2}{\pi}\ E\ I_{cm} = \frac{2}{\pi}\frac{E\ V_{cem}}{R'_L} = \frac{2}{\pi}\frac{E^2}{R'_L}$
- $P_L = 2\frac{1}{T}\int_{t_0}^{t_0+T/2} V_{cem}\ I_{cm}\ \sin^2(\omega t)dt = \frac{V_{cem}^2}{2R'_L} = \frac{E^2}{2R'_L}$
- And consequently, since there is no power dissipated for biasing:
  $P_C = P_D - P_L = \frac{2}{\pi}\frac{E\ V_{cem}}{R'_L} - \frac{V_{cem}^2}{2R'_L}$

We see that P_D increases monotonically with the signal amplitude V_{cem} (there was no dependence in the class A amplifier) - see figure 9.15.

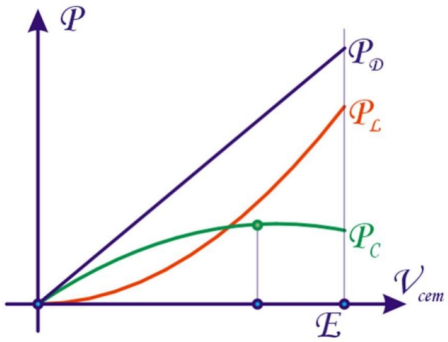


Figure 9.15: Power distribution for class B amplifier

P_C has a quadratic behavior and reaches a maximum when:

$$\frac{dP_C}{dV_{cem}} = \frac{2}{\pi}\frac{E}{R'_L} - \frac{V_{cem}}{R'_L} = 0$$

this is at $V_{cem} = \frac{2}{\pi}E$ and thus

$$P_{C,max} = P_C\big|_{V_{cem}=\frac{2E}{\pi}} = \frac{2E^2}{\pi^2 R'_L}$$

The efficiency η is equal to:

$$\eta = \frac{P_L}{P_D} = \frac{\pi}{4}\frac{V_{cem}}{E}$$

and $\eta_{max} = \frac{\pi}{4}$. In other words, the maximum efficiency, reached when $V_{cem} = V_{cem,max} = E$, is about 78%. The quality factor F is:

$$F = \frac{P_{C,max}}{P_{L,max}} = \frac{2E^2}{\pi^2 R'_L} / \frac{E^2}{2R'_L} = \frac{4}{\pi^2}$$

and for each transistor $F_{Tr} = \frac{2}{\pi^2} \approx 0.2$. So for every 5 W delivered to the load, the transistor consumes only 1 W. This is a lot better then before.

How to generate $-v_i$?

Given v_i, there are basically two ways te generate $-v_i$, as required by the class B amplifier:

1. By using a phase splitter, i.e. an amplifier with $R_C = R_E$, where the outputs are taken at the collector (180° out-of-phase) and at the emitter (in-phase), as in figure 9.16.
2. By using a transformer, as in figure 9.17.

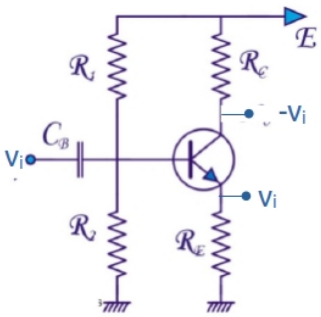


Figure 9.16

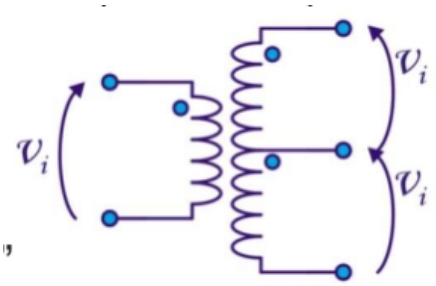


Figure 9.17

What if $V_{BEQ} \neq 0$?

Until now, we have neglected the threshold voltage V_{BEQ} needed to generate a current in the transistor. But in reality, $v_i > 0.6$ V before the transistor conducts. If we take this into account, the resulting current waveform is not a perfect sinusoid, but rather a cascade of half-periods interrupted by regions of zero current, when v_i is too low to generate a forward bias in the base-emitter junctions, as in figure 9.18. This phenomenon is called *cross distortion*.

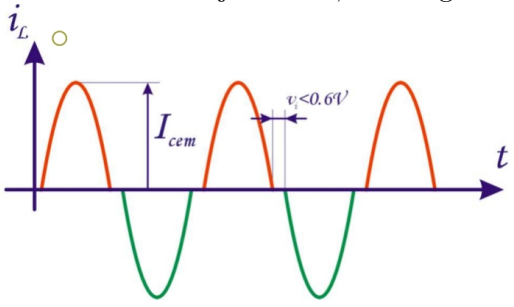


Figure 9.18

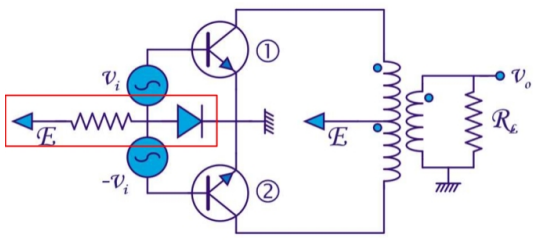


Figure 9.19

A way to solve this is to pre-bias the loop with the base-emitter junction, by adding one threshold voltage via a diode, as in figure 9.19 - the part in red is the pre-bias circuit. This circuit adds one V_{BEQ} to v_i such that a positive v_i is enough to generate a collector current in transistor ① because the voltage at its base will be $v_i + 0.6$ V.

9.4 Push-Pull Amplifiers

A similar idea of using two transistors circuits on top of each other, is the so-called *push-pull* topology of figure 9.20. The goal is not to amplify a voltage, but to deliver current to and from a load R_L.
The circuit consists of an input voltage v_i that will be coupled in through capacitors C_B, two biasing circuits with resistors R_B and R that put the base voltage at $\frac{R_B R}{R_B+R}E$, and two transistors in common-collector configuration (i.e. as emitter followers). Note that transistor ① is an npn, but ② is a pnp transistor. The load R_L is thus connected to the emitters of both transistors.

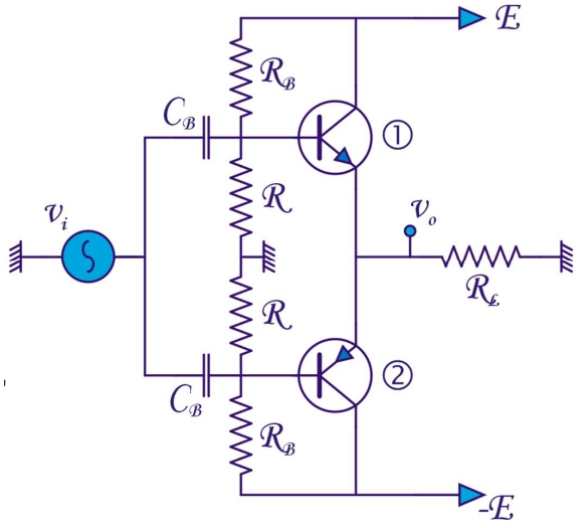


Figure 9.20

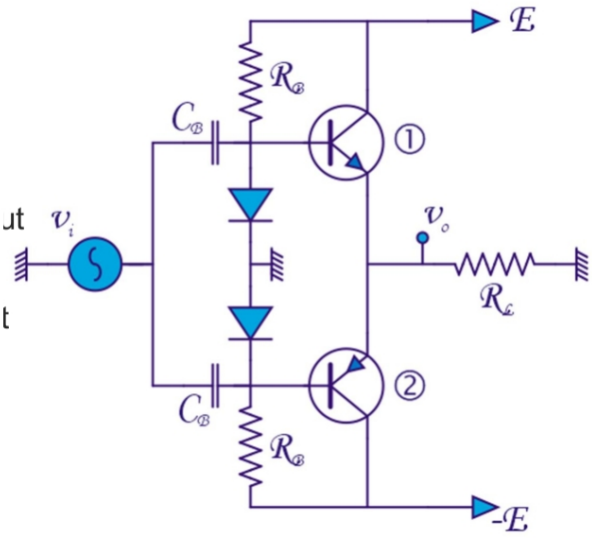


Figure 9.21

If v_i is zero, both transistors carry the same current because of the symmetry, and hence there is no current in the load: $i_{C1} = i_{C2}$ and $i_{RL} = i_{C1} - i_{C2} = 0$. If v_i increases, i_{C1} will increase and i_{C2} will decrease. Consequently, $i_{RL} = i_{C1} - i_{C2} > 0$ and transistor ① will "push" current into the load. On the other hand, if v_i decrease, i_{C2} increases and i_{C1} decreases, and transistor ② will "pull" current out of the load. Because both transistors are emitter followers, $v_o \approx v_i$ and $A_v \approx 1$. To reiterate: the goal is to deliver current, not to increase the gain.
The configuration in figure 9.20 is a class A amplifier because it has a bias circuit with two resistors. If we replace resistors R with diodes as in figure 9.21, we have the same pre-biasing circuitry as the class B amplifier in figure 9.19. In this case, resistor R_B biases the diode to prevent cross distortion.
A way to provide both voltage amplification and current, is by using this circuit as the output trap after an OPAMP as in figure 9.22.

The voltage gain is set by the OPAMP: $A_v = 1 + \frac{R_2}{R_1}$. Note that the feedback via R_2 is applied to the output of the circuit, and not to the output of the OPAMP. Because of its high

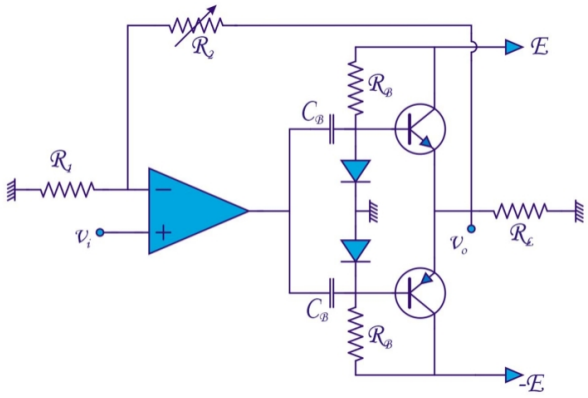


Figure 9.22: Push-Pull amplifier with OPAMP

gain, the OPAMP will try to keep the votage v between its two terminals equal to zero, so that we can write at the negative terminal:

$$v_i = \frac{0/R_1 + v_o/R_2}{1/R_1 + 1/R_2} = \frac{R_1}{R_1 + R_2} v_o$$

and this equation will set the output voltage. In this way, we avoid voltage loss through the push-pull stage.

9.5 Class C Amplifier

For the class C amplifier, we return to the circuit of the class A amplifier in figure 9.6, but we now apply a negative bias (notice the orientation of E_B). This means that the transistor will conduct for only a fraction 2τ of the period T, as the waveform of i_C in figure 9.24 shows.

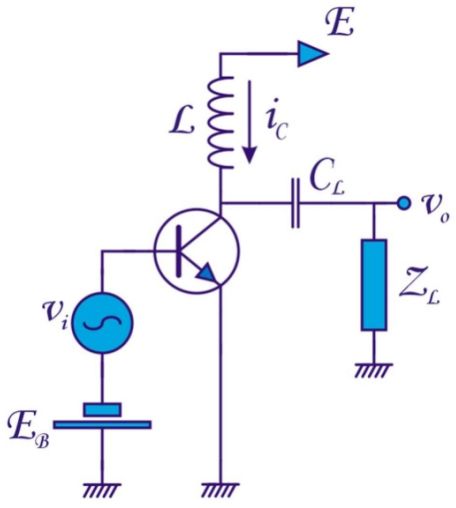


Figure 9.23

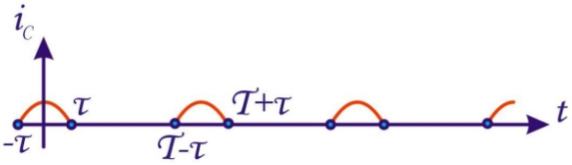


Figure 9.24

The current i_C in figure 9.24 can be expressed as a cosine from which we subtract the lower $\cos(\omega_0\tau)$ part:

$$i_c = I_{cm} \frac{\cos(\omega_0 t) - \cos(\omega_0 \tau)}{1 - \cos(\omega_0 t)} \text{ if } kT - \tau < t < kT + \tau$$

and 0 elsewhere. Because this function is periodic, we can write it as a Fourier series i.e. the spectrum of the function:

$$i_c(t) = I_0 + \sum_{n=1}^{\infty} I_n \cos(n\,\omega_0 t) \text{ with } I_n = \frac{2}{T}\int_{-T/2}^{T/2} i_c(t)\ \cos(n\omega_0 t)\ dt$$

This spectrum is shown in figure 9.25. From this spectrum, we only need the first harmonic, centered on $\omega = \omega_0$.

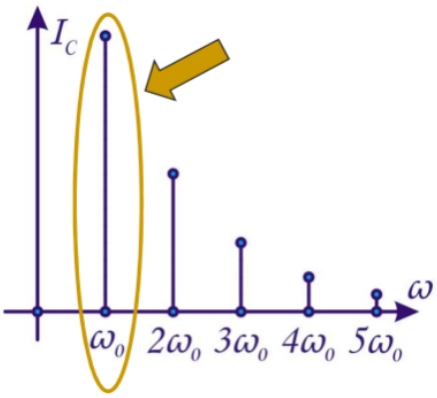


Figure 9.25: Spectrum of figure 9.24

The output voltage v_o is equal to $v_o = Z_L\ i_c$. To just keep the first harmonic, we need a Z_L that is maximum at frequency $\omega_0/2\pi$, and zero at all other frequencies. This can be achieved by using an RLC tank as in figure 9.26 for Z_L. The impedance is:

$$Z(\omega) = \frac{R}{1 + jQ\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right)}$$

with resonance pulsation $\omega_0 = \sqrt{1/LC}$ and quality factor $Q = \omega_0 RC = \omega_0/BW$ as in figure 9.27. So a proper choice of L and C determines ω_0, and the choice of R and C sets the bandwidth. The bandwidth should be set by considering the bandwidth of the signals of interest, and by taking into account that non-linear distortion creates other harmonics.
How the expressions for the RLC circuit are found is explained in section 9.5.1

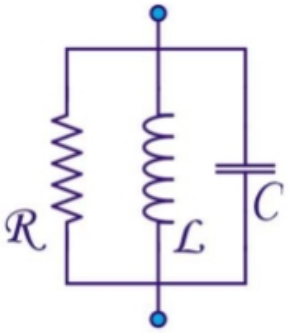


Figure 9.26

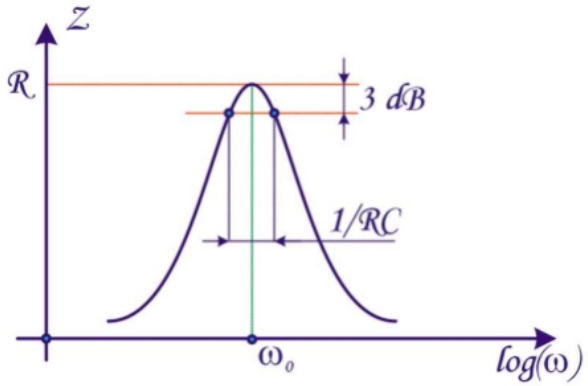


Figure 9.27

The class C amplifier has less dissipation in the transistor and higher power transfer. It is mostly used in telecommunication applications and high power transmission.

9.5.1 The RLC Tank

Consider the parallel combination of a resistance R, a capacitance C and an inductor L, as in figure 9.26. Qualitatively, we see that if $\omega \ll \omega_0$, the inductance will act as a short circuit, and $Z \approx 0$. Similarly, if $\omega \gg \omega_0$, C becomes a short and again $Z \approx 0$. Now let's analyze this circuit more formally.
Working with the admittances, we can write:

$$\begin{aligned} Y &= \frac{1}{R} + j\omega C + \frac{1}{j\omega L} \\ &= \frac{j\omega L + R - \omega^2 RLC}{j\omega RL} \end{aligned}$$

and thus, for the impedance Z, where we use the resonance frequency $\omega_0^2 = \frac{1}{LC}$:

$$\begin{aligned} Z(\omega) &= \frac{j\omega LR}{j\omega L + R - \omega^2 RLC} = \frac{R}{1 - j\frac{R}{\omega L} + j\omega RC} \\ &= \frac{R}{1 + j(\omega RC - \frac{R}{\omega L})} = \frac{R}{1 + jRC(\omega - \frac{1}{\omega LC})} \\ &= \frac{R}{1 + jRC(\omega - \frac{\omega_0^2}{\omega})} = \frac{R}{1 + j\omega_0 RC\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right)} \end{aligned}$$

If $\omega \ll \omega_0$ or if $\omega \gg \omega_0$, $Z \to 0$. If $\omega = \omega_0$, $Z(\omega_0) = R$. The impedance as function of frequency is plotted in figure 9.27. If we define the width where $Z = \frac{R}{\sqrt{2}}$ (i.e. a decrease of -3 dB) as the impedance bandwidth $\Delta\omega$, we can show that $\Delta\omega = \frac{1}{RC}$. This means that the resonance pulsation ω_0 is controlled by L and C, and the bandwidth is set by R and C.
The quality factor is defined as the ratio between ω_0 and the bandwidth:

$$Q = \frac{\omega_0}{\Delta\omega} = \omega_0\ RC$$

With this expression, we can rewrite the expression for Z:

$$Z(\omega) = \frac{R}{1 + jQ\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right)} \tag{9.3}$$

9.6 Class D Amplifier

Figure 9.28 shows the circuit of a class D amplifier. This amplifier uses switches: when the top switch is open, the one in the bottom is closed, and vice versa. This can be implemented with two MOS transistors, as in figure 9.29: a PMOS at the top and an NMOS at the bottom. when the input is high, the v_{GS} of the bottom transistor is high and it will conduct and pull the output to the ground. When the input is low, the top transistor will conduct and pull the output to the supply E. These circuits will be studied in more detail in chapter 13.

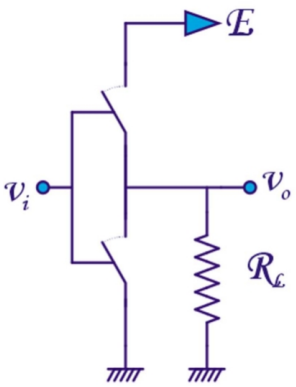


Figure 9.28

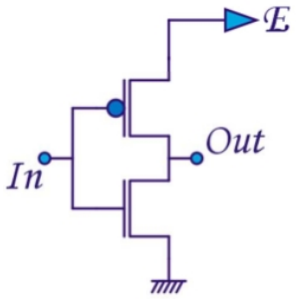


Figure 9.29

The advantage of a switch is that it never dissipates power: one of the switches will always be open, so no bias current can flow. In reality however, each of these switches (transistors) will have a parasitic resistor: each transistor will consume - when closed - a power P_C because it will operate in the linear region (see 5.2.3) and we can model it as a (small) resistor r_{ds} in series with R_L:

$$P_C = i_{ds}\, v_{ds} = \frac{E}{r_{ds}+R_L}\frac{r_{ds}E}{r_{ds}+R_L} = E^2\frac{r_{ds}}{(r_{ds}+R_L)^2} \approx E^2\frac{r_{ds}}{R_L^2}$$

The issue with this amplifier is that the amplitude is not preserved, so we can only transmit signals where the information is not encoded in the amplitude, but for example in the pulse width (pulse width modulation - PWM) or in the frequency (frequency modulation - FM).
If the load is not a resistor but a pure capacitor, i.e. if we replace R_L by C_L, we can compute the associated power dissipation. When the top switch closes and the bottom one is open, a charge $Q = C_L E$ is transferred from the supply to the capacitance C_L. When the bottom switch is closed, the capacitor will discharge and the same amount of charge is transferred to ground. If this happens with a frequency f, we generate an effective current i_{cap} equal to fC_LE. The average power consumption over one period is thus:

$$P_{cap} = i_{cap}\, E = fC_LE^2$$

9.7 Class S Amplifier

A class D amplifier can not be used when the signal amplitude is important, as in AM modulation. However, we can modify a class D amplifier to obtain a class S amplifier that is capable of amplifying AM signals.

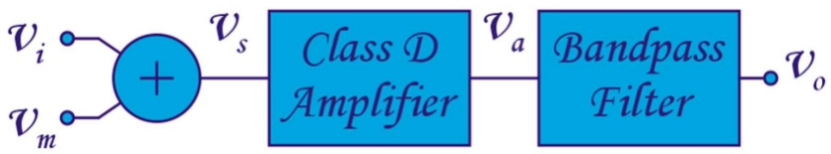


Figure 9.30: Block diagram of a class S amplifier

To do this, we first take the sum of the input signal v_i, which we assume has the form $v_i = A(t)\cos(\omega_0 t)$, with a signal $v_m = \cos(\frac{\omega_0}{2}t)$. We assume that $|A(t)| \ll 1$, so that $|v_m| \gg |v_i|$. By representing these signal as Fresnel vectors as in figure 9.23, with v_m in red, v_i in green and $v_s = v_i + v_m$ in yellow, we see the small green arrow v_i rotates fast around the slowly rotating red arrow v_m. By tracing out the end-points of the green arrow v_i, we see that the amplitude information $A(t)$ of v_i is new encoded in the phase of v_s. We can now use a class D amplifier to amplify the signal, losing all amplitude information but keeping the phase information intact, and extract the amplitude information from the output signal with a bandpass filter, as in figure 9.30.

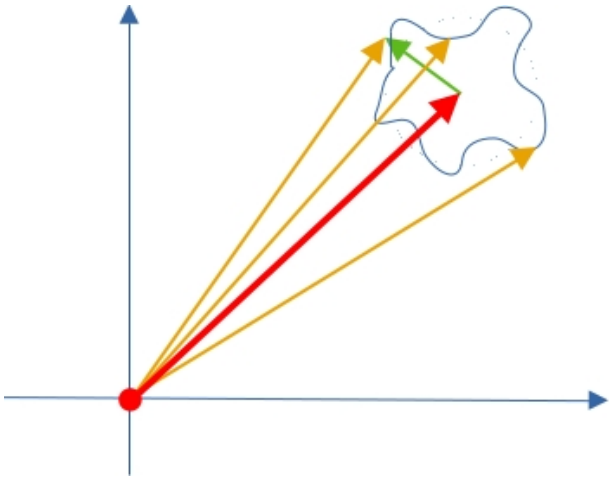


Figure 9.31

Mathematically, we can write:

$$\begin{aligned} v_s &= A(t)\cos(\omega_0 t) + \cos(\frac{\omega_0}{2}t) = A(t)\cos(\frac{\omega_0}{2}t)\cos(\frac{\omega_0}{2}t) - A(t)\sin(\frac{\omega_0}{2}t)\sin(\frac{\omega_0}{2}t) + \cos(\frac{\omega_0}{2}t) \\ &= [1 + A(t)\cos(\frac{\omega_0}{2}t)]\cos(\frac{\omega_0}{2}t) - [A(t)\sin(\frac{\omega_0}{2}t)]\sin(\frac{\omega_0}{2}t) \\ &= \sqrt{[1 + A(t)\cos(\frac{\omega_0}{2}t)]]^2 + [A(t)\sin(\frac{\omega_0}{2}t)]^2}\cos[\frac{\omega_0}{2}t + \arctan\frac{A(t)\sin(\frac{\omega_0}{2}t)}{1 + A(t)\cos(\frac{\omega}{2}t)}] \end{aligned}$$

Because we lose the amplitude, the output of the class D amplifier v_a is equal to:

$$v_a = K\cos[\frac{\omega_0}{2}t + \arctan\frac{A(t)\sin(\frac{\omega_0}{2}t)}{1 + A(t)\cos(\frac{\omega}{2}t)}]$$

with K the gain of the amplifier. When we take into account that $|A(t)| \ll 1$, we can simplify this expression:

$$\begin{aligned} v_a &= K\cos[\frac{\omega_0}{2}t + \arctan\frac{A(t)\sin(\frac{\omega_0}{2}t)}{1 + A(t)\cos(\frac{\omega}{2}t)}] \approx K\cos[\frac{\omega_0}{2}t + \arctan A(t)\sin(\frac{\omega_0}{2}t)] \\ &\approx K\cos[\frac{\omega_0}{2}t + A(t)\sin(\frac{\omega_0}{2})] = K\cos(\frac{\omega_0}{2}t)\cos[A(t)\sin(\frac{\omega_0}{2}t)] - K\sin(\frac{\omega_0}{2}t)\sin[A(t)\sin(\frac{\omega_0}{2}t)] \\ &\approx K\cos(\frac{\omega_0}{2}t) - KA(t)\sin^2(\frac{\omega_0 t}{2}t) \\ &\approx K\cos(\frac{\omega_0}{2}t) - \frac{K}{2}A(t) + \frac{K}{2}A(t)\cos(\omega_0 t) \end{aligned}$$

After the bandpass filter, centered on ω_0, the output signal becomes:

$$v_o \approx \frac{K}{2} A(t) \cos(\omega_0 t)$$

We have thus amplified with high efficiency, and conserved the amplitude $|A(t)|$.

9.8 The Selective Amplifier

Until now, we have amplified a relatively large bandwidth, and we have made no effort to restrict the amplification only to a specific frequency. In most telecommunications applications however, we are often only interested in a specific segment of the spectrum. This means we want to amplify a narrow bandwidth around the center frequency ω_0, and discard all other frequencies.
A perfect solution for this would be to use the RLC-circuit from section 9.5.1, which has as impedance the resistor R at resonance and zero impedance elsewhere as in figure 9.32. However, in reality every inductance L also has a small series resistance r_s. We want to replace this series combination of an ideal L with a parasitic r_s with a parallel combination of L with a resistance R_p as indicated in figure 9.33.

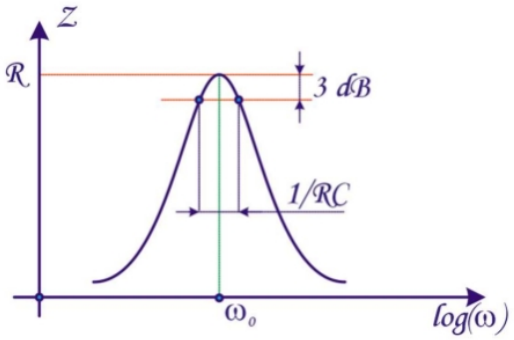


Figure 9.32

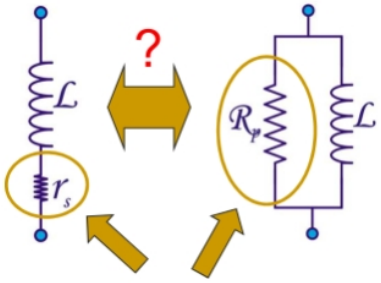


Figure 9.33

The impedance of an inductor L in series with a resistor r_s is:

$$\begin{aligned} Z_{L+r_s} &= j\omega L + r_s \\ \frac{1}{j\omega L + r_s} &= \frac{r_s - j\omega L}{r_s^2 + \omega^2 L^2} \\ &\approx \frac{r_s}{\omega^2 L^2} + \frac{1}{j\omega L} \end{aligned}$$

This last simplification is only valid if $\omega L \gg r_s$, or in other words, if the quality factor of the coil (the inductor) $Q_C = \frac{\omega L}{r_s}$ is a lot larger than 1. This means that we have an equivalent impedance consisting of a resistor $R_p = \frac{\omega^2 L^2}{r_s}$ in parallel with an inductance $j\omega L$.
When we add a capacitor in parallel with R_p and L, we get a true RLC-tank as in figure 9.26 - but in reality we have a non-ideal coil in parallel with C. The quality factor of the whole is $Q = \omega R_p C = \frac{R_p}{\omega_0 L} = Q_C$. So any LC-combination has in reality an impedance as in figure 9.32, but with a maximum of $R_p = \frac{\omega^2 L^2}{r_s}$ and a bandwidth of $\frac{1}{R_p C}$.
In a real RLC-circuit, we can push the resistance r_s to become a resistance R_p in parallel with

the true resistance R. The quality factor if the entire circuit is $Q = \omega_0(R_p||R)C < \omega_0 R_p C = Q_C$. So the use of an external resistor decreases the quality factor.

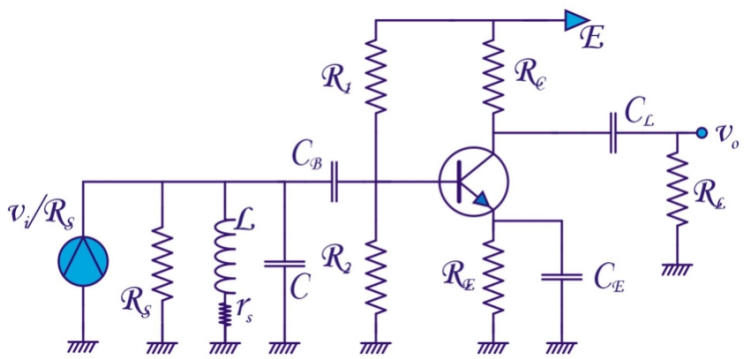


Figure 9.34

To construct a selective amplifier, we add the LC-circuit to the input stage of a common-emitter amplifier, as in figure 9.34. Note how the input has been replaced by the Norton equivalent, namely $Z_N = R_S$ and $I_N = v_i/R_S$.
To compute the central frequency and gain, we introduce the small-signal equivalent circuit in figure 9.35, where we suppose that both C_B and C_E are short circuits, i.e. we are in the normal operating domain, and C_L is also a short circuit.

We can simplify the circuit in figure 9.35 by replacing C_μ by the Miller capacitance $C_M = (1 + gR_{eq})C_\mu$, by introducing an equivalent resistance $R_{eq} = r_c||R_C||R_L$ and by grouping all resistors at the left side together:

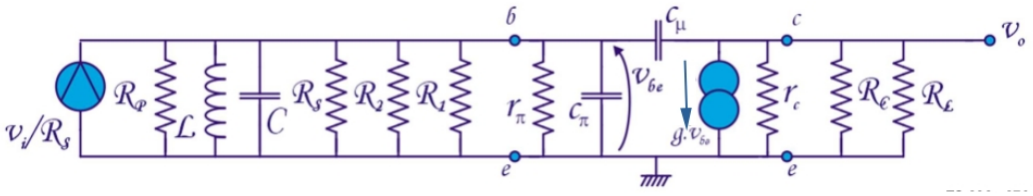


Figure 9.35: The selective amplifier

$$R_\pi = R_p||R_s||R_1||R2||r_\pi$$

The resulting circuit is shown in figure 9.36. The RLC tank is formed by R_π, C_π and L, so the central frequency ω_0 is $\frac{1}{\sqrt{LC_\pi}}$ and the bandwidth $\Delta\omega = \frac{1}{R_\pi C_\pi}$. At ω_0, the impedance at the input is $Z = R_\pi$.
To compute the gain, we observe that $v_{be} = \frac{v_i}{R_S} R_\pi$ - if R_S is very small, this reduces to $v_{be} \approx v_i$ - and the output voltage is:

$$v_o = -R_{eq}\, g\, \frac{v_i}{R_S} R_\pi$$

This is only valid at $\omega = \omega_0$ and so the gain is $A_v = -g\frac{R_\pi}{R_S} R_{eq}$. At frequencies $\omega < \omega_0 - \Delta\omega$ or $\omega > \omega_0 + \Delta\omega$ the gain is (almost) zero.

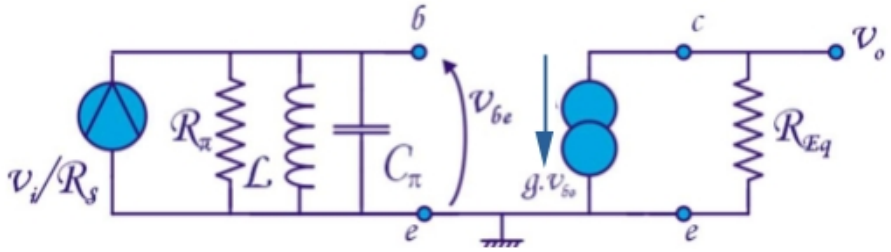


Figure 9.36

Chapter 10

Feedback Theory

When we talk about feedback, we mean that part of the output signal is fed back to the input. What is applied to the input of the circuit is not the input signal, but rather the difference between input signal and the returned output signal.
We will see in this chapter that feedback offers many advantages:

- It allows for accurate gain control,
- You can control the input and output impedance,
- You can extend the bandwidth,
- The distortions are reduced

We discuss these topics in turn, and afterwards have a look at some stability issues that may arise by applying feedback.

10.1 Accurate Gain Control

In figure 10.1, the output voltage v_o, generated by an amplifier with transmittance $A(j\omega)$, is returned to the input through a system with transmittance $H(j\omega)$. We can write:

$$\begin{aligned} v_f &= H(j\omega)\, v_o \\ v_c &= v_i - v_f = v_i - H(j\omega)\, v_o \\ v_o &= A(j\omega)\, v_c = A(j\omega)\, (v_i - H(j\omega)\, v_o) \\ \Rightarrow A_{CL} = \frac{v_o}{v_i} &= \frac{A(j\omega)}{1 + A(j\omega)\, H(j\omega)} \\ &= \frac{A}{1+AH} = \frac{A}{1+T} \end{aligned} \tag{10.1}$$

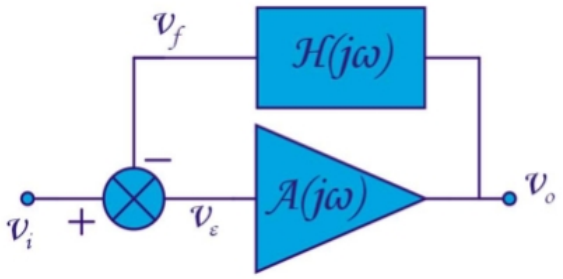


Figure 10.1

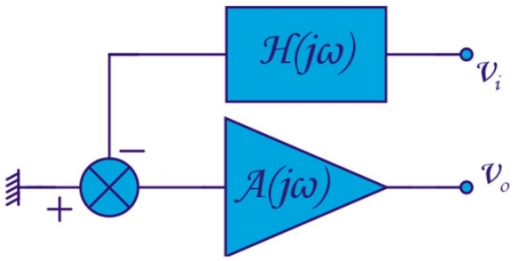


Figure 10.2

Some definitions: A is the open-loop gain and H the feedback gain, A_{CL} is called the closed-loop gain and $T = AH$ is the *loop gain*. The loop gain T can be obtained by this procedure:

1. Open the feedback loop at the amplifier output,
2. Ground the input,
3. Apply v_i at the open end of the loop,
4. Obtain T by computing v_o/v_i.

This method is visualized in figure 10.2, where we applied it to the circuit in 10.1. The signal at the input of the amplifier is $H\ v_i$, and $v_o = AH\ v_i$, so we find the loop gain.

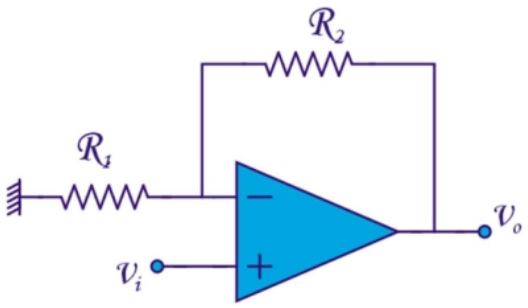


Figure 10.3

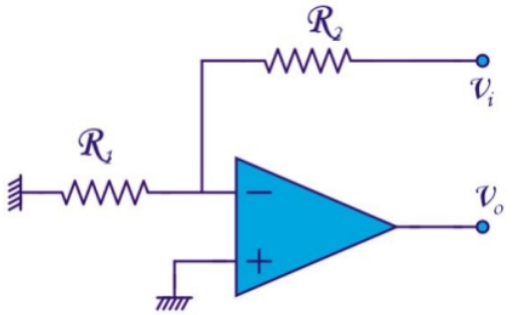


Figure 10.4

When we apply this to the non-inverting OPAMP topology that we saw in section 7.4, which is pictured in 10.3, we get the circuit in figure 10.4 with the feedback loop opened and the input grounded. The voltage at the negative terminal is $v^- = \frac{R_1}{R_1+R_2} v_i$, and the output is then $v_o = -A\ \frac{R_1}{R_1+R_2}\ v_i$. The loop gain T is thus :

$$T = -A\ \frac{R_1}{R_1 + R_2}$$

Typically, the amplifier gain A will be very high. In that case, the closed-loop gain doesn't depend on A:

$$\begin{aligned} A_{CL} &= \frac{A}{1+AH} = \frac{1}{H}\frac{A}{1/H + A} \\ &= \frac{1}{H}\frac{1}{1+T^{-1}} = \frac{1}{H}\frac{T}{T+1} \\ &\approx \frac{1}{H} \end{aligned}$$

Usually, A is not only very large, but also difficult to control. It also depends on the small-signal parameters like g, g_m, r_c, r_{ds}, ... On the other hand, $\frac{1}{H}$ is not very high, but typically only depends on passive elements like resistors and capacitors, which can be very well controlled.
Applying this to the example of the non-inverting amplifier, we had $T = AH = -A\,\frac{R_1}{R_1+R_2}$, so $H = \frac{R_1}{R_1+R_2}$ or $H^{-1} = \frac{R_1+R_2}{R_1} = 1 + \frac{R_2}{R_1}$, just as we found in section 7.4.

10.2 Input and output impedance with feedback

Consider the circuit in figure 10.5. We compute the in- and output impedances Z_i and Z_o taking into account that no current enters the OPAMP, with a negative gain $-A$:

$$Z_i = \frac{v_i}{i_i} = \frac{v_i}{(v_i - v_o)/Z} = \frac{v_i Z}{v_i + A v_i}$$
$$Z_o = \frac{v_o}{i_o} = \frac{v_o}{(v_o - v_i)/Z} = \frac{v_o Z}{v_o + v_o/A}$$

or, in other words:

$$Z_i = \frac{Z}{1+A}$$
$$Z_o = \frac{Z}{1+A^{-1}}$$

This is the same result as we found for the Miller capacitor in section 8.3: there we had $Z = \frac{1}{j\omega C_\mu}$ and the gain was $-gR_{eq}$, so we replaced the input impedance with:

$$Z_i = \frac{Z}{1+A} = \frac{1/j\omega C_\mu}{1+gR_{eq}} = \frac{1}{(1+gR_{eq})j\omega C_\mu} = \frac{1}{j\omega C_M}$$

and hence $C_M = (1+gR_{eq})C_\mu$, as we found previously. The Miller conditions guarantee that the contribution of C_μ at the output can be neglected.

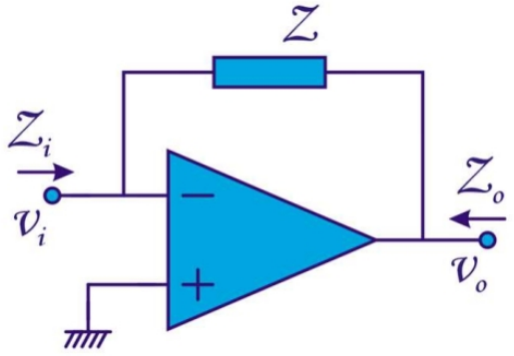


Figure 10.5

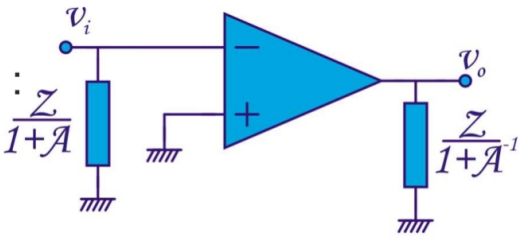


Figure 10.6

By using these relations, we can redraw the circuit as the one in figure 10.6, where we explicitly represented the in- and output impedances.

10.3 Increased bandwidth

Assume that the amplifier $A(j\omega)$ can be modeled as a system with a single pole in ω_0:

$$A(j\omega) = \frac{A_0}{1 + j\frac{\omega}{\omega_0}}$$

Substituting this in the expression for the closed-loop gain A_{CL} gives:

$$\begin{aligned} A_{CL} &= \frac{A(j\omega)}{1 + A(j\omega)H} = \frac{\frac{A_0}{1+j\frac{\omega}{\omega_0}}}{1 + \frac{A_0}{1+j\frac{\omega}{\omega_0}}H} \\ &= \frac{\omega_0 A_0}{(\omega_0 + j\omega) + \omega_0 A_0 H} = \frac{\omega_0 A_0}{(1+T_0)\omega_0 + j\omega} \\ &= \frac{\frac{A_0}{1+T_0}}{1 + j\frac{\omega}{\omega_0(1+T_0)}} \end{aligned}$$

So the closed-loop system is also a first-order system, but now:

- The gain A_0 is divided by the loop gain plus one.
- The bandwidth is increased by the same factor because the pole lies in $\omega_0(1 + T_0)$.
- The gain-bandwidth product GBW remains constant.

This is the same conclusion as in figure 7.33.

10.4 Distortion reduction

We model a distortion as an unwanted signal somewhere downstream in out circuit. Take for example the circuit in figure 10.7, where the first amplifier is an OPAMP, and the second is a class B push-pull amplifier, like in figure 9.22, including the feedback path through R_1 and R_2, modeled with $H(j\omega)$. The second amplifier suffers from a large distortion, modeled as a voltage v_d applied at its input. Without feedback, this distortion is directly visible at the

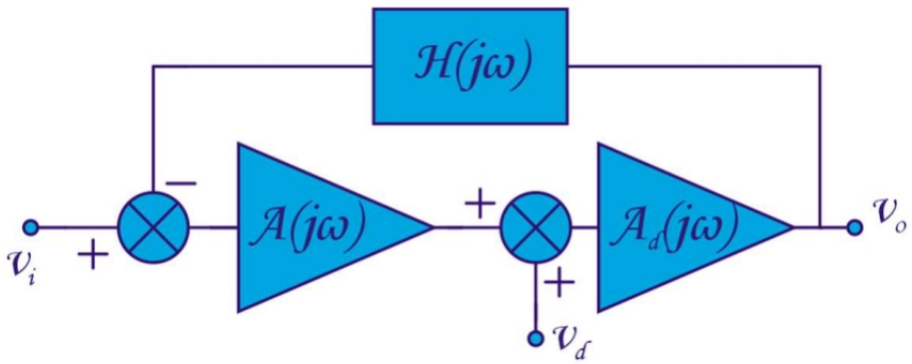


Figure 10.7

output, magnified by the gain A_d : $v_o = A(j\omega)A_d(j\omega)\ v_i + A_d(j\omega)v_d$. However, with feedback this becomes:

$$\begin{aligned} v_o &= A_d(v_d + A(v_i - Hv_o)) \\ (1 + A_dAH)\ v_o &= A_dv_d + A_dAv_i \\ v_o &= \frac{AA_d}{1+T}v_i + \frac{A_d}{1+T}v_d \\ &\approx \frac{1}{H}v_i + \frac{1}{AH}v_d \end{aligned}$$

with the loop gain $T = AA_dH$. This means that the distortion is reduced by a factor A compared to the input signal. Knowing that a distortion downstream will be reduced by the gain of the previous stages, it is important to place a distortion-less amplifier as the first stage of the loop.
In the circuit in figure 9.22, we needed diodes as a pre-bias to avoid cross-distortion. But thanks to the feedback, if you would omit the diodes in the second stage of figure 10.7, there would be no distortion either, because every distortion in the second stage will be divided by A.

10.5 Stability Issues

To keep the operating point of the OPAMP with feedback stable, it is essential that we apply the feedback to the negative terminal. Otherwise, the system becomes unstable and the output will go to infinity (will explode) - in reality, off course, the output will be set to the OPAMP supply $\pm E$.
In principle, we can analyze the frequency response of the feedback topology in figure 10.1 by expressing both the Laplace transforms on $A(s)$ and $H(s)$ as the ratio of two polynomials[1]:

$$A(s) = \frac{N_A(s)}{D_A(s)} \text{ and } H(s) = \frac{N_H(s)}{D_H(s)}$$

and substituting in the expression for the closed-loop 10.1:

$$\begin{aligned} A_{CL}(s) = \frac{A(s)}{1 + A(s)H(s)} &= \frac{\frac{N_A(s)}{D_A(s)}}{1 + \frac{N_A(s)}{D_A(s)}\frac{N_H(s)}{D_H(s)}} \\ &= \frac{N_A(s)D_H(s)}{D_A(s)D_H(s) + N_A(s)N_H(s)} \end{aligned}$$

To verify the stability, we must calculate the poles and check wether their real part is negative (i.e. do they lie in the left half of the s-plane) - see section 2.3.2. If this is the case for all poles, the system is asymptotically stable. However, to find the poles, we must solve $D_A(s)D_H(s) + N_A(s)N_H(s) = 0$. This method is usually very complex because it requires the factoring of high-order polynomials, and it gives no indication of the stability margins, i.e. how much can the different parameters vary before the system becomes unstable.
Let's take another approach. We will assume that the open loop gain $A(j\omega)$ is low-pass: i.e. it has a pole in $\omega = \omega_d$ and for all frequencies $\omega < \omega_d$, $A(j\omega) = A_0$ is constant. For all $\omega > \omega_d$

[1]Note that $s = \sigma + j\omega$. We usually set $\sigma = 0$, i.e. we analyze the frequency response of the system.

the gain decreases. This means we model $A(j\omega)$ with following expression, where ω_d is the dominant pole and ω_{nd} is the non-dominant pole [2]:

$$A(j\omega) = \frac{A_0}{(1 + j\frac{\omega}{\omega_d})(1 + j\frac{\omega}{\omega_{nd}})}$$

With $T(j\omega) = A(j\omega)H(j\omega)$, we know that the closed-loop gain is

$$A_{CL} = \frac{A(j\omega)}{1 + T(j\omega)}$$

This means that if $T(j\omega) > 0$ the system is always stable. The problem arises when $T(j\omega) = -1$, i.e. if for a certain ω_ϕ:

- $|T(\omega_\phi) = 1$, and
- $\angle T(\omega_\phi)) = 180°$

So even when we apply feedback to the negative terminal, if the frequency response is such that the signal undergoes an additional phase shift of 180°, it can create an unstable circuit. The magnitude at the frequency ω where this phase shift happens must be ≥ 1 for this unstable behavior to be sustained.
To demonstrate this, consider the topology in figure 10.8. We established earlier that $T = \frac{R_1}{R_1+R_2}A(j\omega)$. Notice that the feedback is applied to the negative terminal. A small perturbation that occurs at the negative input, as shown by the up arrow, will be amplified and becomes a larger perturbation at the output, but in the other direction. This output perturbation will be transmitted to the input by the feedback path but, as long as $T > 0$, **will act against the original perturbation**. The perturbation can not be sustained and will die out. The circuit is unconditionally stable.
Now assume the feedback is done on the positive terminal as in figure 10.9. The perturbation will travel around the circuit and will arrive back with the same direction at the input node. This means that the perturbation will not only be sustained but magnified, and the system becomes unstable. Note that this situation would be identical to the one in figure 10.8 if T were negative.

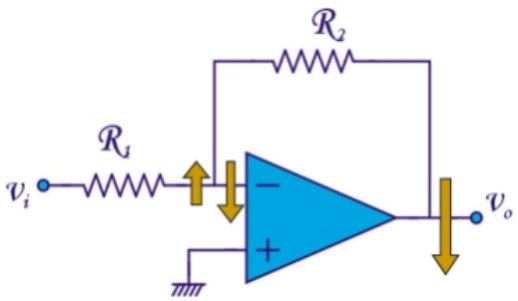


Figure 10.8

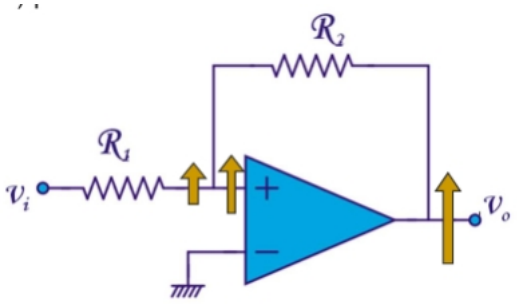


Figure 10.9

[2] In general, there can be many of those non-dominating poles; what is important is that the ω_d determines the low-frequency behavior.

If H has no frequency dependence and $A(j\omega)$ has a single pole ω_d, we can solve the problem analytically. With $A(j\omega) = \frac{A_0}{1+j\frac{\omega}{\omega_d}}$, we find for the (closed) loop gain:

$$T(j\omega) = \frac{A_0 H}{1 + j\frac{\omega}{\omega_d}} \text{ and } A_{CL} = \frac{A_0}{1 + A_0 H} \frac{1}{1 + j\frac{\omega}{\omega_d(1+A_0 H)}} = \frac{A_{CL0}}{1 + j\frac{\omega}{\omega_{CL}}}$$

which is a result we already found when we discussed the increased bandwidth. Interpreting this as the movement of pole in the s-plane, we see that the existing pole ω_d has moved to the left:

$$\omega_{CL} = (1 + A_0 H)\, \omega_d$$

This movement is shown in figure 10.10. This is the *root-locus*, because it traces how the root(s) (poles) of the closed-loop response move as function of H. The closed-loop always remains stable. This can also been learned from the Bode curve, because a first-order system can only have a maximal phase shift of 90°, and not 180° as needed for an instability.

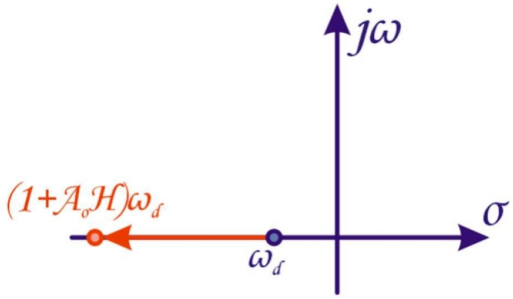


Figure 10.10

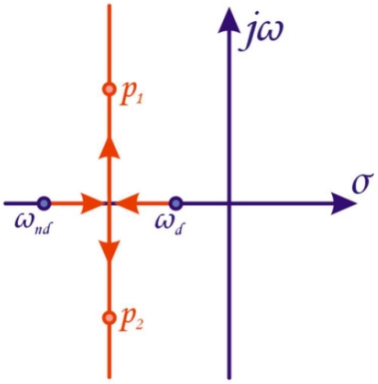


Figure 10.11

We can do the same analysis for a system with two poles, where $A(j\omega)$ is given by:

$$A(j\omega) = \frac{A_0}{(1 + \frac{\omega}{\omega_d})(1 + \frac{\omega}{\omega_{nd}})}$$

with ω_d the dominant and ω_{nd} the non-dominant pole, as before. Then:

$$T(j\omega) = A(j\omega)H(j\omega) = \frac{A_0 H}{(1 + \frac{\omega}{\omega_d})(1 + \frac{\omega}{\omega_{nd}})}$$

and

$$A_{CL} = \frac{A(j\omega)}{1 + T(j\omega)} = \frac{A_0 \omega_d \omega_{nd}}{(1 + A_o H)\omega_d \omega_{nd} + j\omega(\omega_d + \omega_{nd}) - \omega^2}$$

The poles of A_{CL} are found by solving for the roots of the denominator. The result is:

$$p_{1,2} = -\frac{1}{2}(\omega_d + \omega_{nd}) \pm \frac{1}{2}\sqrt{(\omega_d + \omega_{nd})^2 - 4(1 + A_0 H)\omega_d \omega_{nd}}$$

This means that if $1 + A_0 H$ is small, the poles are real and located close to $-\omega_d$ and $-\omega_{nd}$. As $A_0 H$ increases, they move closer together. At some point (i.e. when the part under the square root is zero) they collide and are both equal to $-\frac{1}{2}(\omega_d + \omega_{nd})$. If $A_0 H$ increases further,

the term under the square root becomes negative and the poles acquire an imaginary part, i.e. they become complex when

$$A_0 H > \frac{(\omega_d - \omega_{nd})^2}{4\omega_d\,\omega_{nd}} \approx \frac{\omega_{nd}}{4\omega_d}$$

This situation is represented in the root-locus of figure 10.11. Note how even for a second-order system, the poles always remain in the left half-plane and so the system stays stable. The imaginary component of the poles is reflected in the fact that the circuit may oscillate temporarily and may appear unstable.
The maximal phase-shift of a second order system is 180°, but this shift only occurs when $\omega \to \infty$ so in practice the system will remain stable. This however does not mean that there are no issues. Even stable systems can have a large overshoot (i.e. how far goes the step response beyond its final value) in the step response, as in figure 10.12. This figure shows the step response of a second order system. In general, the transfer function of a second order system can be written as:

$$H(j\omega) = \frac{\omega_0^2}{\omega_0^2 + 2j\zeta\omega_0\omega - \omega^2} \tag{10.2}$$

with ω_0 the eigenfrequency and ζ the damping ratio. The damping ratio ζ is related to the ratio of the imaginary and real component of the complex pole: if the poles are both real, $\zeta > 1$ and there are no oscillations; if $\zeta = 0$, the poles are imaginary and the system is undamped: even a small perturbation is enough to generate a continuous oscillation. If ζ is small, the overshoot is high, but the response is fast. Ideally, $\zeta \approx 0.46$ to have a reasonable fast step response without too much overshoot, as can be seen in figure 10.11.

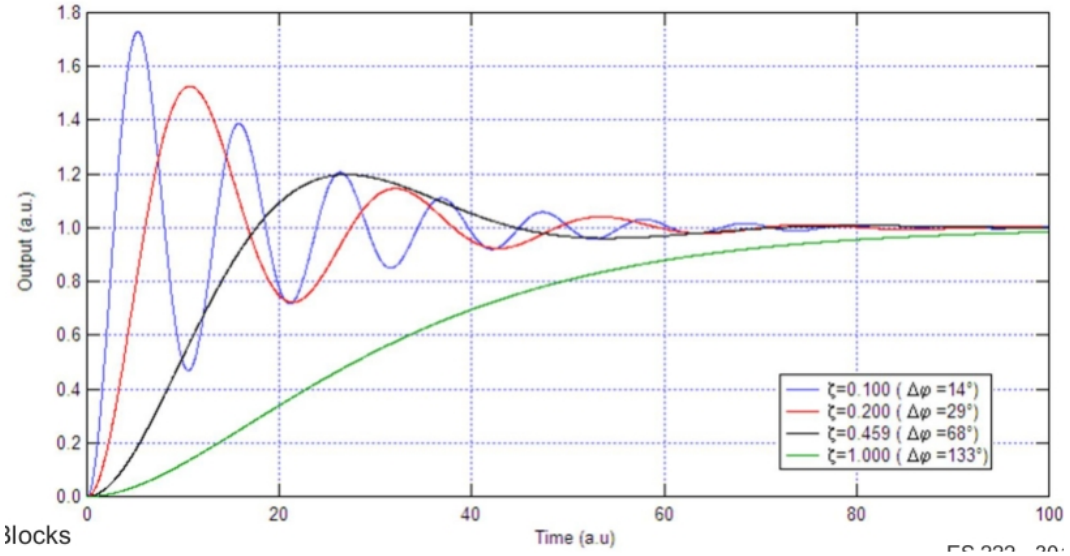


Figure 10.12

The damping ratio is related to the *phase margin* $\Delta\phi$. This is the difference between the phase at the frequency where the gain is 0 dB (i.e. $A_v = 1$) and a phase of 180°. Figure 10.13

shows the (normalized) bode curves and phase margin for different values of ζ. The ideal ζ corresponds to a phase margin $\Delta\phi = 68°$. If $\Delta\phi = 0$, the system is unstable. So in a sense the phase margin gives an indication how far we are from instability.

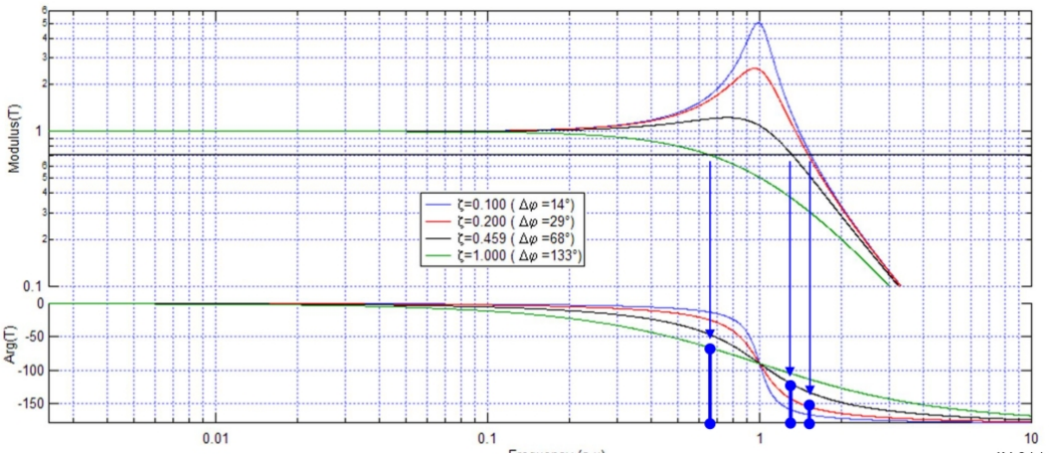


Figure 10.13

In summary we can say that (1) OPAMPs have multiple (2 or more) poles and (2) this will have an effect on circuit stability. This stability depends on the poles (and zeros) of the loop gain $T(j\omega)$ and thus implicitly on $A(j\omega)$ en $H(j\omega)$, of which the latter term is application dependent. In essence, there are two types of OPAMPS:

1. Compensated OPAMPS: for these devices, as long as $|H(j\omega)| < 1$, stability is guaranteed. These circuits have a large phase margin, but are inherently slow because their bandwidth is artificially reduced.

2. Uncompensated OPAMPS: no measures were taken to assure stability; the user is responsible for this. These devices do have maximal bandwidth.

Chapter 11

Oscillators

Usually, what we want to do when applying feedback is to control the loop gain $T(j\omega)$ such that the amplifier becomes and stays stable. However, in this chapter we try to do the opposite: for a certain frequency ω_0, we design the circuit such that $T(j\omega_0) = H(j\omega_0)A(j\omega_0) = 1$. Note how this is different from the condition in chapter 10: we assume here that the feedback signal is *added* to the input signal.
This means that the circuit will be unstable for one single frequency ω_0. These circuits generate sinusoidal output signals at ω_0, even when there is no signal at the input[1]. We call them *oscillators*.

11.1 Phase Shift Oscillator

This oscillator consists of a single common-emitter amplifier that (a) provides the gain and (b) provides a 180° phase shift. To generate the other 180° phase shift, 3 RC blocks are placed in series at the output of the amplifier, as in figure 11.1. So we have an amplifier, followed by feedback loop of 3 RC blocks, that couple the output signal back to the input. If each RC circuit could function independent from the others, we could just tune R and C to provide a 60° shift for a required frequency ω_0. However, because each next block loads the previous one, the calculation is more complicated.
To determine ω_0, we establish the AC equivalent circuit as in figure 11.2 - with $R_{eq} = r_c||R_C$ and $R_\pi = r_\pi||R_1||R_2$ - and compute the loop gain $T(j\omega)$. Then we verify the oscillation conditions:

- $|T(j\omega_0)| = 1$
- $\angle T(j\omega_0) = 0$

so for one frequency, the signal travels around the loop and arrives in phase and with the same magnitude back at the input[2] This condition also implies that $\Im m\{T(j\omega)\} = 0$ and $\Re e\{T(j\omega)\} = 1$.
To compute the loop gain, we cut the link between the input and output of the transistor, and compute which v'_{be} at R_π will be generated by the current source $g\ v_{be}$. The loop gain is

[1] There is always *noise* present, and white noise contains every frequency. This will be studied in other courses

[2] If $|T(j\omega_0)| > 1$, we wont get a nice sinusoid but a block signal at the output.

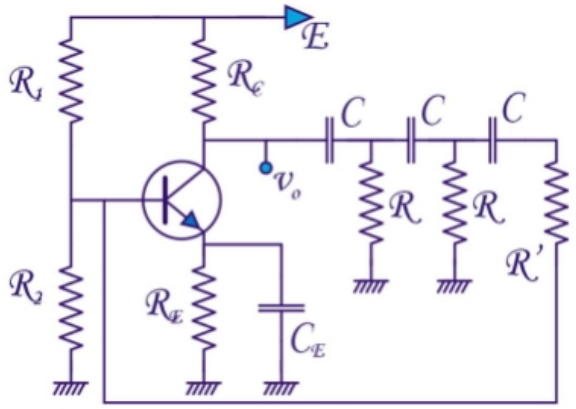


Figure 11.1

then $T = \frac{v'_{be}}{v_{be}}$. The resistance R' is chosen such that $R_\pi + R' = R$.
From the AC equivalent circuit, we find:

$$-gv_{be} = -gR_\pi i_b \approx i_b\left(3 + \frac{4}{j\omega RC} - \frac{1}{\omega^2 R^2 C^2} + \frac{R}{R_{eq}} + \frac{6}{j\omega R_{eq}C} - \frac{5}{\omega^2 C^2 R R_{eq}} - \frac{1}{j\omega^3 R^2 R_{eq} C^3}\right)$$

The imaginary part gives us ω_0:

$$\begin{aligned}
\frac{4}{RC} + \frac{6}{R_{eq}C} - \frac{1}{\omega_0^2 R^2 R_{eq} C^3} &= 0 \\
\Rightarrow \frac{4}{R} + \frac{6}{R_{eq}} &= \frac{1}{\omega_0^2 R^2 R_{eq} C^2} \\
\Rightarrow 4R_{eq} + 6R &= \frac{1}{\omega_0^2 RC^2} \\
\Rightarrow \omega_0^2 &= \frac{1}{(4R_{eq} + 6R)RC^2}
\end{aligned}$$

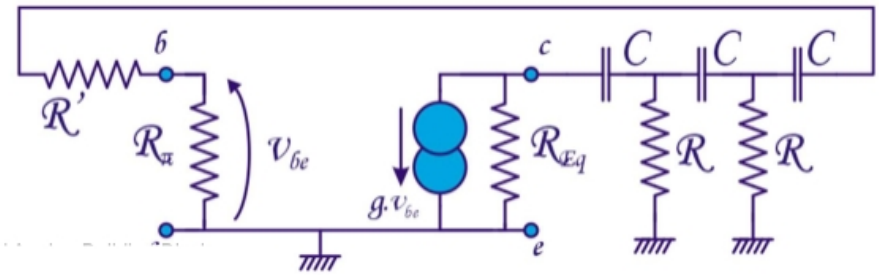


Figure 11.2

The real part provides the criterion for oscillation:

$$\begin{aligned} -gR_\pi &= 3 + \frac{1}{\omega^2 R^2 C^2} + \frac{R}{R_{eq}} - \frac{5}{\omega^2 C^2 R R_{eq}} \\ \Rightarrow -gR_\pi &= 3 - \frac{6R + 4R_{eq}}{R} + \frac{R}{R_{eq}} - \frac{5(6R + 4R_{eq})}{R_{eq}} \\ (gR_\pi - 23) &+ 4\frac{R_{eq}}{R} - 29\frac{R}{R_{eq}} = 0 \end{aligned}$$

This equation can be solved for the ratio $\frac{R}{R_{eq}}$:

$$\frac{R}{R_{eq}} = \frac{gR_\pi - 23}{58} + \sqrt{\frac{(gR_\pi - 23)^2 - 464}{58}}$$

This equation only has a solution only if $gR_\pi > 23 + \sqrt{464} \approx 44.6$, which puts a lower limit on the transconductance g and so on the bias current I_{CQ} of the amplifier. If g is is higher, the output will be distorted; if it is lower, there will be no oscillation.

11.2 Wien Bridge Oscillator

The Wien bridge oscillator from figure 11.3 is a popular oscillator that uses an OPAMP with two feedback loops:

- One stable loop through resistances R_S and R_F
- An unstable loop (on the positive terminal) through C in series with R ($= Z_F$), and C in parallel with R ($= Z_S$):

$$\begin{aligned} Z_F &= R + \frac{1}{j\omega C} = \frac{1 + jRC\omega}{j\omega C} \\ Z_S &= R || \frac{1}{j\omega C} = \frac{R}{1 + jRC\omega} \end{aligned}$$

The loop gain can be found by computing the input of the OPAMP $v_i = v_p - v_n$ and by setting $v_o = Av_i$. This comes down to cutting the loop at the output of the amplifier, injecting a signal v_t into the feedback path and computing v_o:

$$\begin{aligned} v_n &= \frac{R_S}{R_S + R_F} v_o \\ v_p &= \frac{Z_S}{Z_S + Z_F} v_o = \frac{j\omega RC}{1 + j3RC\omega - R^2C^2\omega^2} v_o \\ v_o &= A\, v_i = A\, (v_p - v_n) \\ &= A\left(\frac{j\omega RC}{1 + j3RC\omega - R^2C^2\omega^2} - \frac{R_S}{R_S + R_F} \right) v_o \end{aligned}$$

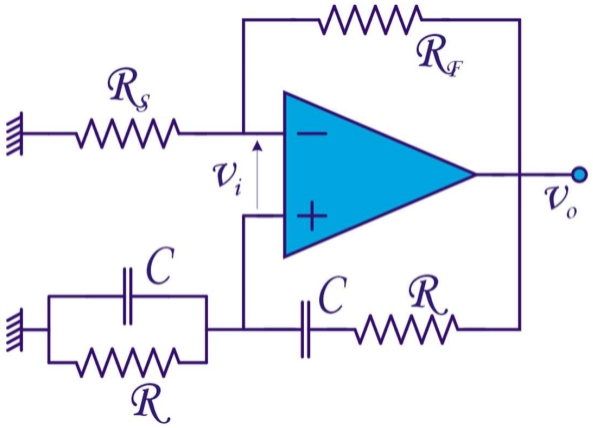


Figure 11.3

With $A \gg 1$, this becomes:

$$\frac{j\omega RC}{1 + j3RC\omega - R^2C^2\omega^2} \approx \frac{R_S}{R_S + R_F} = k$$
$$\Rightarrow j\omega RC = k + j\omega 3kRC - \omega^2 kR^2C^2$$

In this case, the real part gives ω_0:

$$k - \omega_0^2 kR^2C^2 = 0$$
$$\Rightarrow \omega_0 = \frac{1}{RC}$$

and the oscillation condition is given by the imaginary part:

$$j\omega RC = j\omega 3kRC$$
$$\Rightarrow k = \frac{R_S}{R_S + R_F} = \frac{1}{3}$$

11.3 Colpitts Oscillator

A Colpitts oscillator is an oscillator that's very popular in radio application. It consists of a common-emitter amplifier followed by an inductor and two capacitors in the feedback path. Inductor L_C and capacitor C_B are used for respectively biasing and coupling in the signal, as seen previously. The same goes for capacitor C_E, which short-circuits R_E. We can thus omit them from the AC equivalent circuit in figure 11.5, where we do consider the parasitic transistor capacitances c_π and c_μ and replace the latter by the Miller capacitance C_M (i.e. we suppose that the Miller conditions are satisfied).

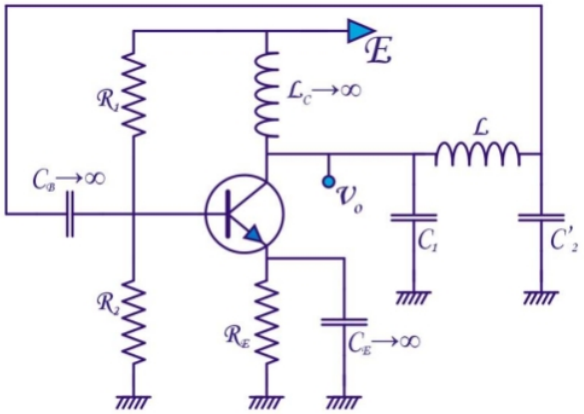


Figure 11.4

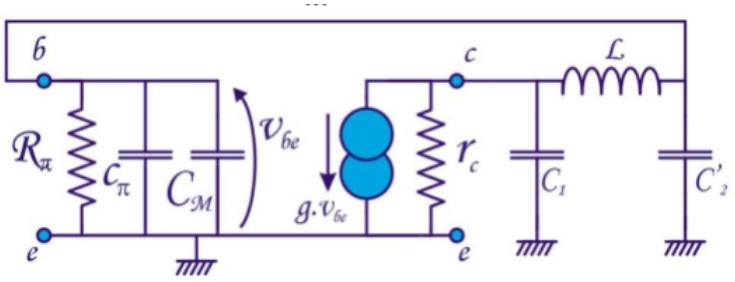


Figure 11.5

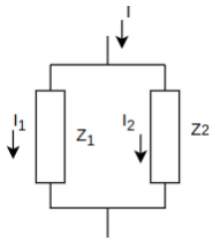


Figure 11.6

We can simplify further by grouping all capacitors between base and ground: $C_2 = c_\pi + C_M + C'_2$. Let $Z_1 = r_c||C_1 = \frac{r_c}{1+j\omega r_c C_1}$ and $Z_2 = j\omega L + (\frac{1}{j\omega C_2}||R_\pi) = j\omega L + \frac{R_\pi}{1+j\omega R_\pi C_2}$. As such, Z_1 and Z_2 are two impedances in parallel and they operate as a current divider (see figure 11.6):

$$I_2 = \frac{Z_1}{Z_1 + Z_2} I$$

Like for the phase-shift oscillator, we assume that the base-emitter voltage v_{be} is generated with the current source $-g\ v_{be}$. The current through the parallel combination of R_π and C_2 is then:

$$i_{R_\pi||C_2} = -g\ v_{be} \frac{Z_1}{Z_1 + Z2}$$

and v_{be} is equal to:

$$\begin{aligned} v_{be} &= \frac{R_\pi}{1 + j\omega R_\pi C_2} i_{R_\pi||C_2} = -g\ v_{be} \frac{Z_1}{Z_1 + Z_2} \frac{R_\pi}{1 + j\omega R_\pi C_2} \\ &= -g\ v_{be} \frac{\frac{r_c}{1+j\omega r_c C_1}}{\frac{r_c}{1+j\omega r_c C_1} + j\omega L + \frac{R_\pi}{1+j\omega R_\pi C_2}} \frac{R_\pi}{1 + j\omega R_\pi C_2} \\ &= -v_{be} \frac{g r_c R_\pi}{j\omega L(1 + j\omega R_\pi C_2)(1 + j\omega r_c C_1) + R_\pi(1 + j\omega r_c C_1) + r_c(1 + j\omega R_\pi C_2)} \end{aligned}$$

This last expression results in the condition:

$$j\omega L - \omega^2(R_\pi C_2 + r_c C1)L - j\omega^3 r_c R_\pi C_1 C_2 L + R_\pi + j\omega r_c R_\pi C_1 + r_c + j\omega R_\pi r_c C_2 = -g r_c R_\pi$$

By examining the real parts of this equation, we find ω_0:

$$\omega_0^2 = \frac{L + r_c R_\pi (C_1 + C_2)}{r_c R_\pi C_1 C_2 L} = \frac{1}{r_c R_\pi C_1 C_2} + \frac{C_1 + C_2}{C_1 C_2 L} \approx \frac{C_1 + C_2}{C_1 C_2 L} = \frac{1}{C_{eq} L}$$

with C_{eq} the parallel combination of C_1 and C_2.
From the imaginary part, the oscillation condition on the transconductance g can be found:

$$g + \frac{R_\pi + r_c}{r_c R_\pi} \approx \frac{C_1 + C_2}{C_1 C_2 L}(R_\pi C_2 + r_c C_1)$$

11.4 Quartz Oscillator

Regular oscillators are not very good time references: they drift and have a precision of only $\sim 1\%$. This means that daily they have an error of about 15 minutes. A *quartz oscillator* - i.e. an oscillator based on the vibrations of quartz crystal - has a much higher precision: from $\sim 10^{-6}$ (1 ppm - about 30 seconds per year) to $\sim 10^{-7}$ if they are kept at a stabilized temperature.
Quartz is a mineral composed of silicon and oxygen atoms arranged in a specific crystal structure. It exhibits a *piezoelectric effect*, which means that it can generate an electric charge when subjected to mechanical stress, such as pressure or vibration. This effect also works in the other way: quartz changes shape when a voltage is applied.

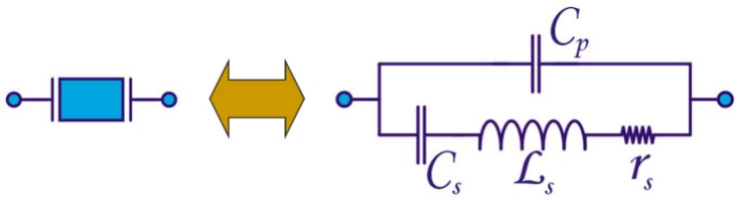


Figure 11.7

The quartz crystal with the external clamps is represented by the symbol in the left part of figure 11.7. From an electrical point of view, a quartz crystal can be modeled as in the circuit on the right part of the figure. The different components are justified because the moving crystal has several mechanical characteristics for which we use an electrical equivalent:

- Stiffness, because you have to provide energy before anything moves: modeled by a capacitor $C_s \sim 10$ fF
- Mass, because there is inertia when you have to displace mass: modeled by an inductor $L_s \sim 1000$ H
- Friction, because heat is generated when the crystal moves: modeled by a resistance $r_s \sim 100\ \Omega$
- The clamps (due to the external connectors): modeled by a a capacitor $C_p \sim 10$ pF

If we ignore r_s, we can compute the admittance $Y(j\omega)$ of the quartz crystal:

$$Y(j\omega) = j\omega C_p + \frac{1}{j\omega L_s + \frac{1}{j\omega C_s}} = j\omega C_p + \frac{j\omega C_s}{1 - \frac{\omega^2}{\omega_s^2}}$$
$$= j\omega C_p \frac{\left(1 + \frac{C_s}{C_p}\right)\omega_s^2 - \omega^2}{\omega_s^2 - \omega^2} = j\omega C_p \frac{\omega_p^2 - \omega^2}{\omega_s^2 - \omega^2}$$

with important frequencies $\omega_s^2 = \frac{1}{L_s C_s}$ and $\omega_p^2 = \left(1 + \frac{C_s}{C_p}\right)\omega_s^2$.
This admittance is purely imaginary. When we plot the imaginary part of the impedance $Z(j\omega) = \frac{1}{Y(j\omega)}$ as function of ω, we find the curve in blue in figure 11.8, with an asymptote in ω_p. The curve in red is the same result, but now with the friction resistance r_c. If $\omega < \omega_s$ or $\omega > \omega_p$, the imaginary part of Z is negative, so Z is capacitive. When $\omega_s < \omega < \omega_p$, Z becomes inductive.

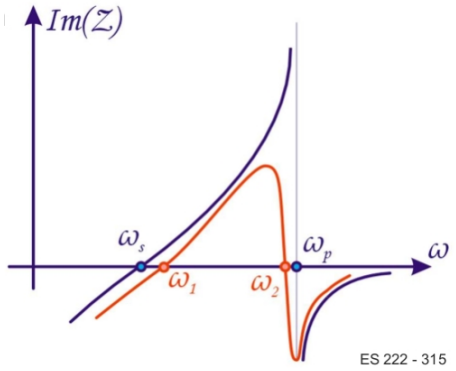


Figure 11.8

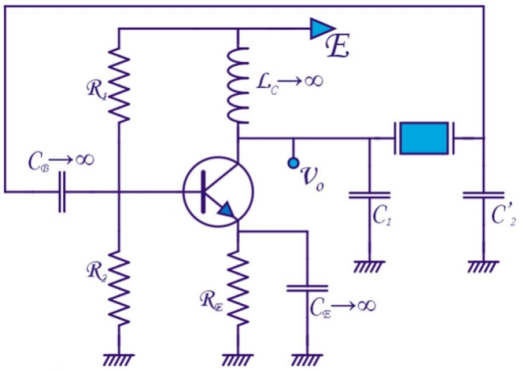


Figure 11.9

We can construct a Colpitts oscillator with a quartz crystal instead of an inductor L, as in figure 11.9. This oscillator has an oscillation frequency $\omega_0 = \frac{1}{LC}$ with C the parallel combination of C_1 and C_2. This frequency necessarily lies between ω_s and ω_p because Z is only inductive in this region, and the corresponding inductance is determined by the red curve in figure 11.8. If ω_0 lies outside this region, Z would be capacitive and the oscillator would not work. If the frequency ω_0 would try to drift, e.g. because of a variation in C, the inductance L would also change, and the positive slope of Z will adjust L so that ω_0 remains relatively constant:

$$C \uparrow \Rightarrow \omega_0 = \sqrt{\frac{1}{LC}} \downarrow \Rightarrow L = Z(j\omega_0) \downarrow \Rightarrow \omega_0 \uparrow$$

This means the system is self-corrective and it is also why the quartz oscillator is very stable.

11.5 Relaxation Oscillator

The oscillators we saw so far were all based on the principle that for a certain frequency ω_0, we want to create a loop gain $T(j\omega_0) = 1$. Another type of oscillator is the *relaxation oscillator*, which is based on the instability of an amplifier. We configure the circuit such that the output of an amplifier keeps switching between high and low, and a capacitor elsewhere in the circuit will eternally be charged and discharged (the *relaxation* of a capacitor). The output waveform will no longer be a sinusoid, but a block or triangle signal.
Whether a circuit is stable or not can be verified by Lyapounov's theorem. To do this:

- First find all the fixed points of the circuit, i.e. the solutions of the governing equations with all time derivatives and inputs set to zero.
- Linearize the system around these solutions.
- Identify how the operating point moves as time t increases (typically, the solutions have an exponential behavior).
- If the operating point returns to the solution, the system is stable. If not, the system might be unstable.

As an example, consider the circuit in figure 11.10, with its operating equations:

$$\begin{aligned} v_o &= \phi(v) \text{ (the OPAMP characteristic)} \\ i &= \frac{v_i + v}{R} = C\,\frac{dv_c}{dt} \\ v_o &= -v + v_c \end{aligned}$$

If $\frac{dv_c}{dt} = 0$, we find the fixed point - with the input set to 0: $i = 0, v_i = 0, v = 0, v_o = 0$.

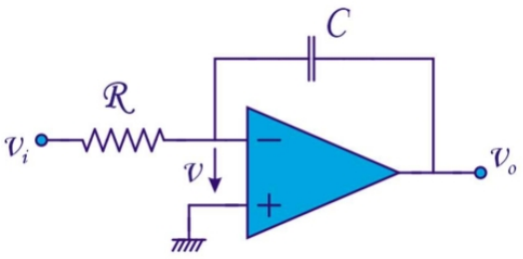


Figure 11.10: An integrator

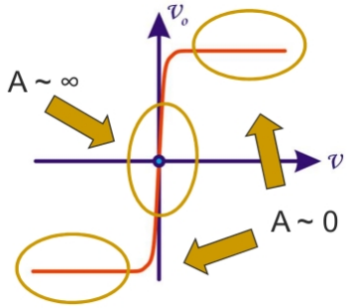


Figure 11.11: OPAMP function: $v_o = \phi(v)$

We linearize the operating equations around this fixed point. To do this, we approximate $v_o = \phi(v)$ by three different linear regions as in figure 11.11: two with gain $A \approx 0$, and one with $A \approx \infty$. The fixed point $v = v_o = 0$ is associated with $A \approx \infty$.

$$\begin{aligned} v_o = A\,v = -v + v_c \Rightarrow v &= \frac{v_c}{A+1} \\ (A+1)v_i = (A+1)RC\,\frac{dv_c}{dt} &- v_c \\ \Rightarrow v_c(t) &= (1+A)v_i(e^{t/(1+A)RC} - 1) \end{aligned}$$

So this system is unstable because $v_c \to \infty$ if $t \to \infty$ (except when $v_i = 0$). This is evident from the exponent in $e^{\alpha t}$: if $\alpha > 0$, the behavior is unstable, as we find here for $\alpha = \frac{1}{(1+A)T}$. The origin of the exponential behavior is a first-order differential equation:

$$\frac{dv_c}{dt} = \alpha\, v_c + v_i$$

which is stable if $\alpha < 0$.
Because $e^x \approx 1 + x$ when x is small, we find that for small values of t, $v_c(t) \approx \frac{t}{RC} v_i$, i.e. we

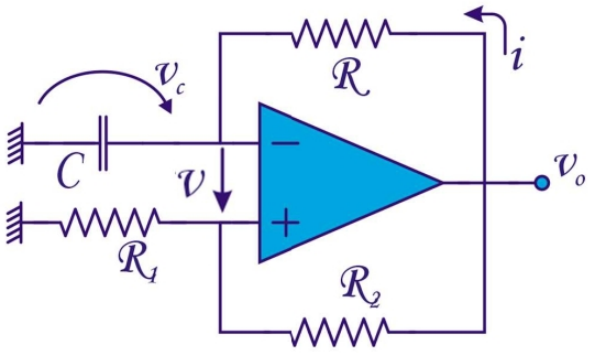


Figure 11.12: A relaxation oscillator

find the expression for an integrator, as expected.

Let's apply this to the circuit in figure 11.12, which is a *relaxation oscillator*.
The operating equations are:

$$\begin{aligned} v_o &= \phi(v) \\ i &= C\,\frac{dv_c}{dt} \\ v_o - v_c &= R\,i \\ v + v_c &= \frac{R_1}{R_1+R_2} v_o = k\,v_o \end{aligned}$$

These equations have a fixed point around $v_o = v_c = v = 0$ and $i = 0$. When we linearize around this point (i.e. we set $v_o = A\,v$), we find:

$$\begin{aligned} v_c = v_o - R\,i = Av - RC\,\frac{dv_c}{dt} \\ v + v_c = kv_o = Ak\,v \\ \rightarrow v_c = (Ak-1)\,v \\ \rightarrow v_c = \frac{A}{Ak-1} v_c - RC\,\frac{dv_c}{dt} \end{aligned}$$

and so:

$$\begin{aligned} v_c &= \frac{Ak-1}{Ak-A-1} RC \frac{dv_c}{dt} \\ &= \frac{k-1/A}{1-k+1/A} RC \frac{dv_c}{dt} \end{aligned}$$

The stability criterion is thus:

$$\frac{k-1/A}{1-k+1/A} > 0$$

or in other words: $A > \frac{1}{k}$.
To understand how the circuit works, assume that $v_o = E$. Then capacitor C is charging

and the voltage at the negative input increases. At some point in time, the voltage at v^- becomes larger than $v^+ = \frac{R_1}{R_1+R_2}E$ and because of the large gain, the output voltage switches to $v_o = -E$. The capacitor C has still a charge of $C\frac{R_1}{R_1+R_2}E$ on it, so v^- changes abruptly to $-E + \frac{R_1}{R_1+R_2}E$. Now, the process reverses: the capacitor starts to discharge, the voltage at v^- decreases and suddenly v becomes positive and the output switches again. The process is shown in figure 11.13.

When C is charging or discharging, we can write:

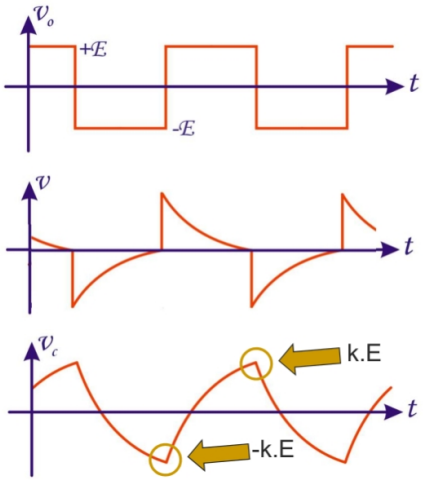


Figure 11.13

$$v_c(t) = v_c(\infty) + (v_c(0) - v_c(\infty))e^{-t/T}$$

With this equation, we can compute the frequency of oscillation: the switching criterion is when $v = 0$, i.e. at time $t = T$ when

$$kE = E + (-kE - E)\, e^{-\frac{T}{2RC}}$$

and thus $T = 2RC \ln \frac{1+k}{1-k}$.

11.6 The Fantastron

The *fantastron* (figure 11.14) is a type of relaxation oscillator that consists of

1. An unstable comparator (also called a *Schmidt trigger*, see also section 15.4): if v_p switches sign, the output voltage will switch from $+E$ to $-E$ because the feedback happens on v^+.

2. An integrator with time constant RC.

With v_a equal to $\pm E$, we can write the voltage v_p as (Millman):

$$v_p = \frac{R_2 v_o \pm R_1 E}{R_1 + R_2}$$

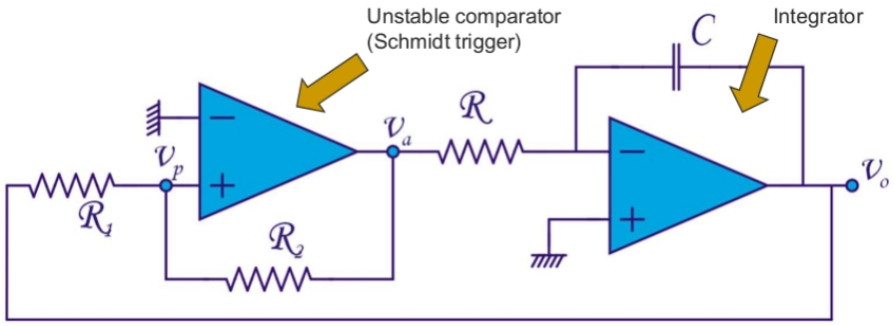


Figure 11.14: The Fantastron

If $v_p > 0$, $v_a = E$, and the integrator will start to integrate down with slope $-\frac{E}{RC}$, until the output reaches $v_o = \pm E\frac{R_1}{R_2}$. At that time, v_p will become negative, v_a switches suddenly to $v_a = -E$, and the process starts all over again in the other direction. The waveforms for v_a and v_o are shown in figure 11.15. In this circuit, we have simultaneously access to a block signal as to a triangle waveform.

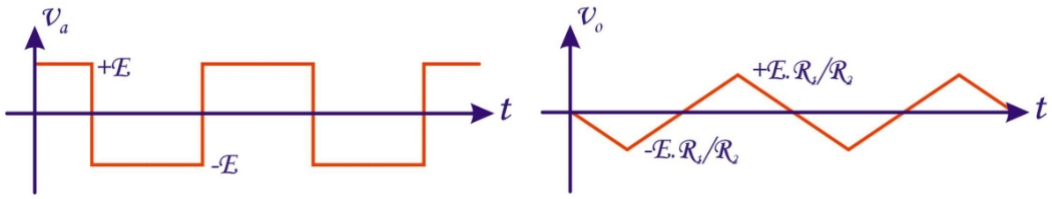


Figure 11.15

The period is thus determined by when v_p switches sign. One half period $\frac{T}{2}$ corresponds to the time for v_o to go from $-\frac{R_1}{R_2}E$ to $+\frac{R_1}{R_2}E$ with slope $\frac{E}{RC}$:

$$-\frac{R_1}{R_2}E + \frac{E}{RC}\frac{T}{2} = \frac{R_1}{R_2}E$$

and so:

$$T = 4\,\frac{R_1}{R_2}\,RC$$

Chapter 12

DC Voltage Generation

The problem we address in this chapter is the generation of a fixed DC voltage of arbitrary value, without too much variation (*ripple*). The supply we have at our disposal is the local AC voltage, which can have a value of 110 V, 210 V or 220 V, and a frequency of 50 Hz or 60 Hz, depending on where you are in the world.

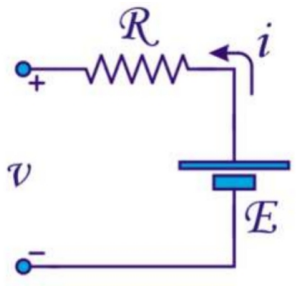


Figure 12.1

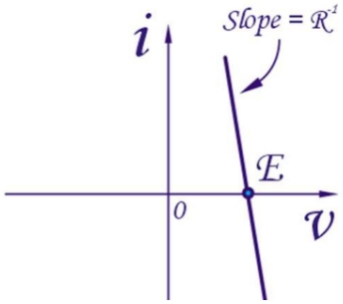


Figure 12.2

An ideal DC voltage source does not exist: every source will have an internal impedance R as in figure 12.1 such that the generated voltage will decrease with the current drawn from the source as in figure 12.2. In an ideal source, the load line would be a vertical line. Furthermore, every source will exhibit a ripple: there will be variations around an average value of the output voltage, even if the load stays the same. Our goal is to implement a DC power supply, generated from an available AC power supply, with minimum ripple e and a low internal impedance.

We will do this in couple of steps. First, the voltage of the AC supply will be converted with a transformer to a level that is more suited. Then, a rectifier, together with a low-pass filter will be used to generate a DC voltage that still has a significant ripple. This ripple will be (partially) removed by using a voltage stabilizer.

12.1 AC voltage modification

Our first task is to modify (reduce) the AC voltage: we try to transfer from 220 V to closer to a voltage that we need (about $\sim$ 10's of Volts). This can easily be done by using a transformer, as in figure 12.3. By setting the turns ratio n, we obtain a smaller amplitude:

$$v_o = \frac{v_i}{n}$$

and the current through the load is $i_o = n\ i_i$, i.e. the power through the load $v_o(t)\ i_o(t) = v_i(t)\ i_i(t)$ is preserved. However, losses can occur and the efficiency decreases for higher powers or currents.

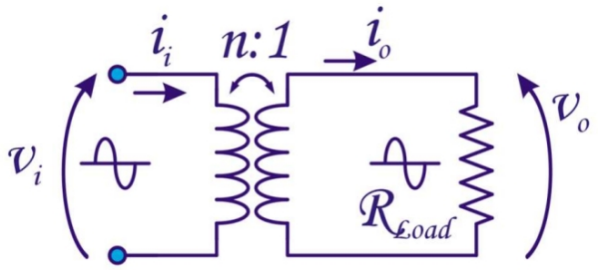


Figure 12.3

12.2 Diode Rectifier

To transform an AC signal to a DC signal, we first try by using a diode as rectifier as in figure 12.4. In this way, there will only be a current i_o when $v_i > 0$. We can compute the average

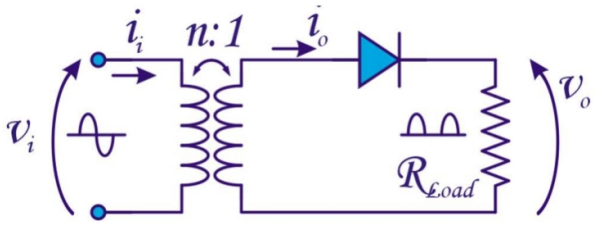


Figure 12.4

output current:

$$I_{CC} = \frac{1}{T}\int_{t=t_0}^{t_0+T} i_{Load}(t)\ dt = \frac{1}{T}\int_{t=t_0}^{t_0+T/2} I_m \sin(\omega t)\ dt = \frac{I_m}{\pi} = \frac{V_m}{\pi\ R_{Load}} \approx 32\%\frac{V_m}{R_{Load}}$$

This means that we lose more than 60% of the output voltage swing. Additionally, we also lose one diode threshold voltage. A potential improvement is to use a germanium diode instead of silicon, because of the lower threshold voltage.
An improvement that allows conduction during both parts of the cycle, is the *Graetz cell* in figure 12.5. During the positive part of the cycle, the two horizontal diodes conduct; during the negative part the diagonal diodes will conduct. So there is a current through the load during the entire cycle and this current always flows in the same direction. We can compute the DC current, as before:

$$I_{CC} = \frac{1}{T}\int_{t=t_0}^{t_0+T} i_{Load}\ dt = \frac{2}{T}\int_{t=t_0}^{t_0+T/2} I_m \sin(\omega t)\ dt = \frac{2I_m}{\pi} = \frac{2V_m}{\pi\ R_{Load}}$$

The signal is fully rectified and the current has doubled compared to the singe diode circuit. Note however that we lose 2 diode threshold voltages because we always have 2 diodes in series with the load.

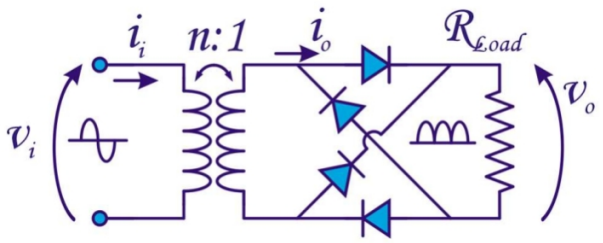


Figure 12.5

We see that the output voltage still has a very high ripple as in figure 12.7, which can be (partially) removed with a low-pass filter. Because the signal has been rectified, the frequency has doubled: if the AC signal is 50 Hz, the cut-off frequency $f_c \approx 100$ Hz. An example is the use of a simple RC-filter as in figure 12.6

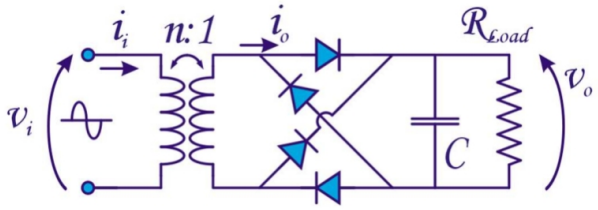


Figure 12.6

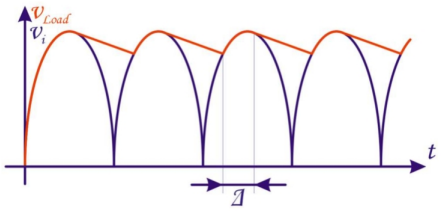


Figure 12.7

During the first part, where v_i is increasing, the capacitor C will charge until v_i reaches a maximum. When v_i will decrease, there is no more current through the top diode because with the capacitor charge, it is reversed biased. During that time, the capacitor will discharge through the load with a time constant $R_{Load}C$ which should be longer than half the period of the AC signal. So v_{Load} decreases because C discharges and v_i increases during the second half period. When $v_{Load} = v_i$, the capacitor will start to charge again until v_i reaches an optimum (a minimum in this case, but because of the rectification through the other diode it is seen by the load as a maximum) and the whole process starts again. The period during which the capacitor charges is called Δ. This is also the only time when the diode conducts. During the other part of the cycle, the capacitor will provide the current to the load.
This also means that when the load R_{Load} receives an average current I_{CC} during a period T, the total displaced charge is $Q_{Load} = I_{CC}T$. But this charge has to be provided by the diode during a time Δ. So:

$$Q_{Load} = I_{CC}T = I_{Diode}\Delta$$

which means that $I_{Diode} = I_{CC}\frac{T}{\Delta}$. Consequently, if we try to make the ripple e small, e.g. by choosing a large C, the charge period Δ will become small and the diode peak current will be very high, which can damage the diode. This means that a ripple can never be completely removed in the rectifier.

12.3 Voltage Stabilizer

To further reduce the ripple, we can use a voltage stabilizer. This circuit will generate a DC voltage based on a reference, typically a Zener diode although sometimes we can use an integrated reference based on more advanced transistor effects.

The principle is shown in figure 12.8. The input is an average voltage V_{IN}, perturbed by a ripple v_{in}. We will always choose the Zener voltage V_Z of the diode such that $V_{IN} + v_{in} > V_Z$. This diode is reversed biased and so the voltage seen by the load is V_Z, i.e we cut everything off above V_Z. Effectively, we remove the difference between v_{IN} and V_Z, i.e. the voltage loss is $v_{IN} - V_Z$. This is the loss over resistor R, such that the power loss through R equals:

$$P_R = \frac{(v_{IN} - V_Z)^2}{R}$$

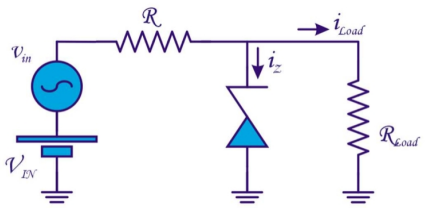


Figure 12.8

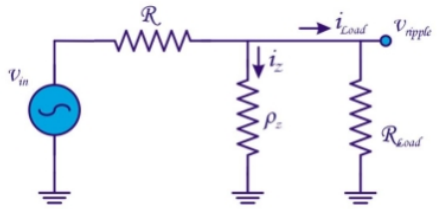


Figure 12.9

Because the power loss is inversely proportional to R, these resistors are normally relatively large and possibly cooled.
The diode load line can be calculated as:

$$v_{IN} - v_Z = R\,(i_{Load} + i_Z) = R\,\left(\frac{v_Z}{R_{Load}} + i_Z\right)$$
$$\Rightarrow \left(1 + \frac{R}{R_{Load}}\right) v_Z = v_{IN} - i_Z R$$

This load in figure line is the green line 12.10, that goes through

$$(v_Z = \frac{R_{Load}}{R_{Load} + R} v_{IN},\ i_Z = 0) \text{ and } (v_Z = 0,\ i_Z = \frac{v_{IN}}{R})$$

The intersection of this line with the red Zener diode characteristic gives the operating point Q. This operating point must lie above the hyperbole of maximum power dissipation for the diode, determined by $V_D\ I_D = P_{max}$. Note that the operating point Q moves when v_{IN} changes: when v_{IN} is too large the diode will burn, when v_{IN} becomes to small we'll leave the Zener region. There are two types of external variations:

- With constant R_{Load}, the current through the load is constant: $i_{Load} = \frac{V_Z}{R_{Load}}$. Any variation in v_{IN} leads to a change in current through R and this change has to be absorbed by the Zener diode.
- With a constant v_{IN}, the current through R is constant: $i_R = \frac{v_{IN} - V_Z}{R}$. If R_{Load} changes, the current through the load changes, and it is again the Zener diode who has to absorb this variation.

These variations can be computed by using the AC-equivalent circuit in figure 12.9. The Zener current varies:

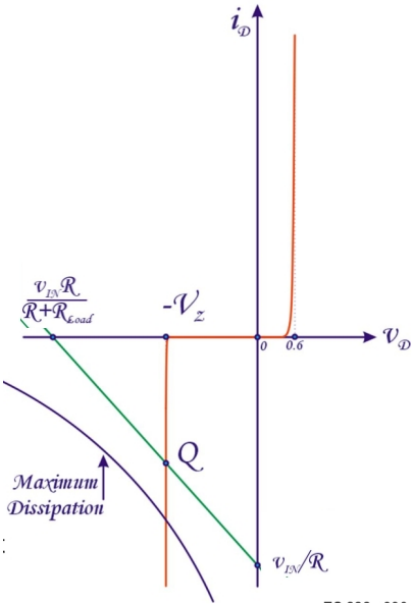


Figure 12.10

- With the input voltage: $i_z = \frac{v_{in}}{R}$,
- With the load voltage (assuming $v_{in} = 0$, i.e. v_{IN} is constant): $i_z = -i_{load}$

This is difficult to control and during normal operation, the diode is likely to exit the Zener domain.
The residual ripple at the output can be obtained with:

$$\frac{v_{ripple}}{v_{in}} = \frac{\rho_z || R_{Load}}{R + \rho_z || R_{Load}} \approx \frac{\rho_z}{R + \rho_z} \approx \frac{\rho_z}{R}$$

and this is small because ρ_z is a very small resistance ($\sim \Omega$). The output impedance is $R_{out} = \rho_z || R_{Load} || R \approx \rho_z$.

12.3.1 Double Stabilizer

To avoid the difficulty with controlling the operating point, we can use the circuit in figure 12.11. This circuit isolates the input voltage v_{in} from the load R_{Load} because the current though resistor R_2 is fixed and determined by the difference in Zener voltages:

$$i_{R2} = \frac{v_{Z1} - v_{Z2}}{R_2}$$

As a consequence, diode D_1 now absorbs the input voltage variation v_{in}: $i_{z1} = \frac{v_{in}}{R_1}$ and diode D_2 absorbs load variations: $i_{z2} = -i_{load}$.
The small-signal circuit of figure 12.12 allows to compute the ripple. By transforming v_{in}, R_1 and ρ_{Z1} to the Thevenin equivalent with $v_{th} = \frac{\rho_{Z1}}{\rho_{Z1}+R_1} v_{in}$ and $Z_{th} = \rho_{Z1} || R_1$, we find:

$$\begin{aligned} \frac{v_{ripple}}{v_{in}} &\approx \frac{\rho_{Z1}}{R_1 + \rho_{Z1}} \frac{\rho_{Z2} || R_{Load}}{\rho_{Z2} || R_{Load} + R_2 + \rho_{Z1} || R_1} \\ &\approx \frac{\rho_{Z1}\rho_{Z2}}{R_1 R_2} \end{aligned}$$

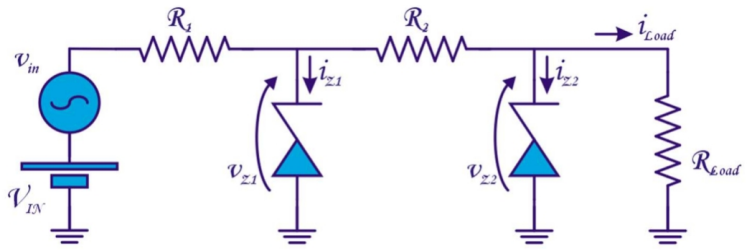


Figure 12.11

So in addition to obtaining a more stable operating point, the ripple is significantly reduced. But note that there is an additional voltage and power loss over resistor R_2, because V_{Z1} has to be larger than V_{Z2}.
Additionally, a good diode can handle currents of a couple mA, but not much more. We need another solution for loads that require more current, or for cases where the load (and hence the required current) is unknown. That's where we can use a transistor-based stabilizer.

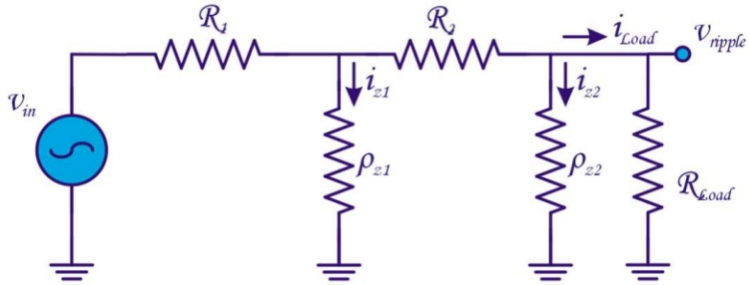


Figure 12.12

12.3.2 Transistor-based Stabilizer

This stabilizer with transistor is an example of a series-stabilizer: the stabilizer is in series with the load. The Zener-diode keeps the base of the transistor at a fixed voltage V_Z, and the voltage at the load is equal to $V_Z - V_{BEQ}$. Hence the current through the load, and consequently also the I_{CQ} of the transistor, is equal to $i_{load} = \frac{v_Z - V_{BEQ}}{R_{Load}}$
The voltage provided to load is set by the Zener-diode with the base-emitter junction of the transistor, i.e. $= V_Z - V_{BEQ}$. The transistor has to able to dissipate a power P_T:

$$P_T = v_{CE}\, i_{load} = (v_{IN} - (v_Z - V_{BEQ}))\frac{v_Z - V_{BEQ}}{R_{Load}}$$

which can be high.
Because the voltage at the transistor base is fixed, the current i_R through resistor R only varies with v_{in}. When i_{Load} remains constant (i.e. the load doesn't change) the base current i_B is also constant and all the variations of i_R are completely absorbed by the diode. So:

$$i_z = \frac{v_{in}}{R}$$

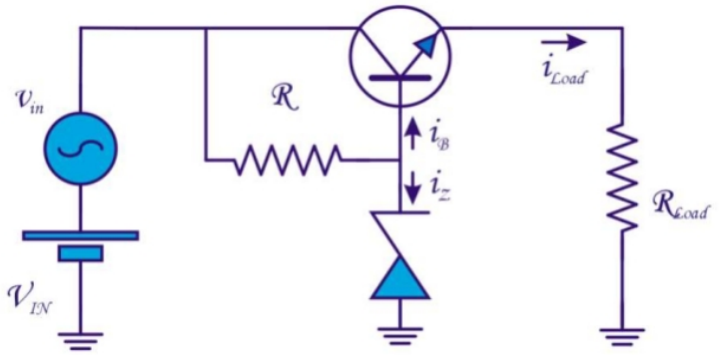


Figure 12.13

The base current i_B only depends on i_{Load} : $i_B = i_{Load}/\beta$. So when v_{in} remains constant and there is no current variation through R, i_z has to vary with i_{Load}:

$$i_z = -\frac{i_{Load}}{\beta}$$

Once again, i_z depends on both v_{in} and R_{Load}, which is not ideal. However, because of the factor β, the variations due to the load are under control. Note how current variations of a couple of ampères are possible.
To compute the ripple and output impedance, we should draw the AC equivalent circuit. But we can redraw the circuit as in figure 12.14 and note that this is a common-collector amplifier. The output impedance is obtained by looking into the emitter of the transistor, so $R_{out} \approx \frac{1}{g}$.

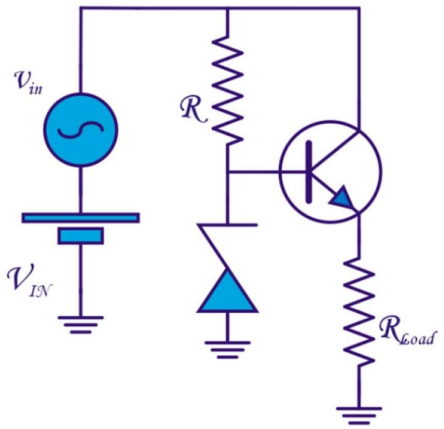


Figure 12.14

For the ripple, we see that the output at the emitter follows the base voltage variations, and this variation in turn is, for small signals, determined by the voltage divider formed by R and ρ_Z:

$$\frac{v_{ripple}}{v_{in}} = \frac{\rho_Z}{\rho_Z + R} \approx \frac{\rho_Z}{R}$$

just as we found for the stabilizer with a single Zener-diode.
One way to isolate the reference diode from the load is with an operational amplifier in

feedback, as in figure 12.15. The OPAMP will force the output voltage to be equal to voltage at its positive node, i.e. V_Z. For a fixed v_{IN}, the voltage drop across resistor R is fixed and equal to $v_{IN} - V_Z$, and so is the current through R. So any variation in the load has no impact on i_Z. The reference only suffers from a dependence of variations of v_{IN}: $i_z = \frac{v_{in}}{R}$. The ripple is like before determined by the voltage divider R - ρ_Z: $\frac{v_{ripple}}{v_{in}} = \approx \frac{\rho_Z}{R}$. The output

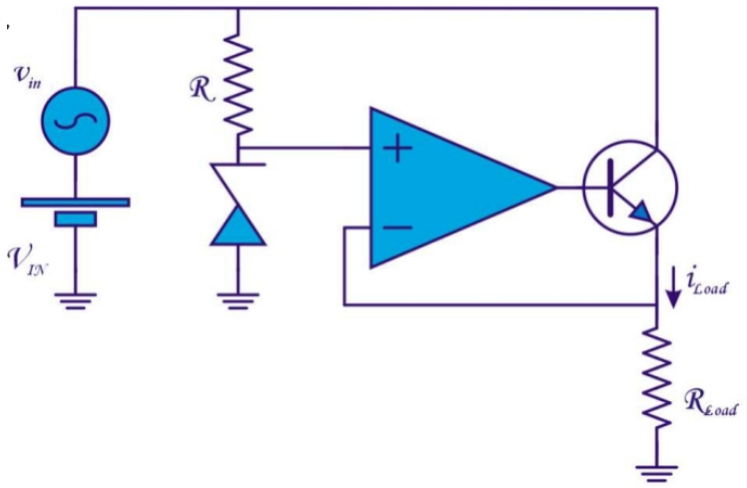


Figure 12.15

impedance as seen by the load is $R_{out} \approx \frac{1}{A_V g}$. This is because the impedance looking into the emitter is $\frac{1}{g}$, and feedback reduces the output impedance by a factor of about A_v (see section 10.2). The output impedance is thus extremely low.

12.4 Supply Protection

The goal of a supply protection is to avoid that a load will draw too much current from the DC supply: this could cause damage to the circuitry. We could use a fuse, but this is in general slow, and it might happen that part of the circuit already has been damaged before the fuse has molten. So we typically rely on electronic circuitry that acts fast and saves the circuit. There are two types of protection:

- A *current limiter*, as in figure 12.16. If the current exceeds a certain threshold i_{max}, the circuit shuts down: there is still current, but the supplied voltage becomes zero.

- A *current fall-back* circuit (figure 12.17) that shuts both current and supply voltage down and reduces them to zero. This is much better than a current limiter, but requires a reset to restore proper operation.

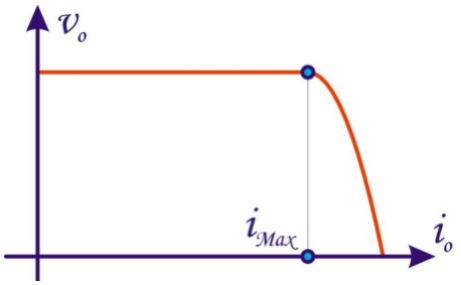


Figure 12.16

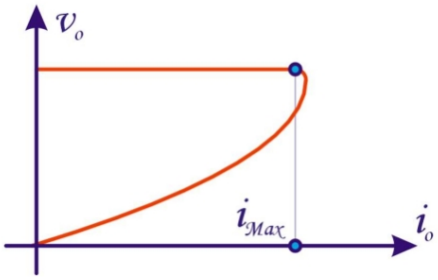


Figure 12.17

We will examine how a current limiter can be implemented. Consider the circuit in figure 12.18. This is the improved output stage of figure 12.15. The voltage V_{ref} is generated by a Zener diode. The resistor R is chosen such that $I_{max} = \frac{V_{BEQ}}{R}$. For a small load current, the bottom transistor is blocked and the top transistor provides the current, just as in figure 12.15. The OPAMP sets the output current to V_{ref}.

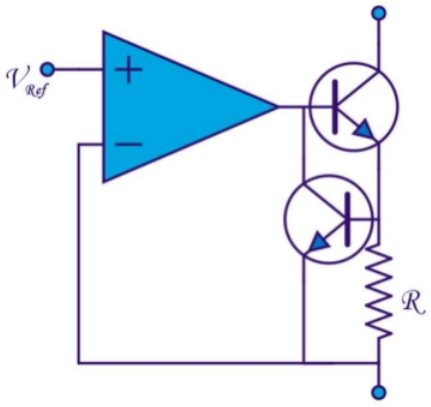


Figure 12.18

As the current increases, so will the voltage drop over R. When the current reaches I_{max}, the bottom transistor has enough base-emitter voltage and it starts to conduct. It will steal base current from the top transistor and any additional current can no longer be provided by this transistor. If there is current beyond I_{max}, it has to come from the OPAMP. But the OPAMP is a voltage amplifier, and is not equipped to provide much current. The OPAMP will no longer be able to supply the required current and breaks down, which causes the output voltage to drop to zero.

12.5 Switched Supply

PM

Part III

Digital Electronics

Chapter 13

Basics of Digital Circuits

Until now, we considered a transistor mostly as an analog element: by applying a voltage at the input (e.g. the base or gate), we generate a continuous and varying current which in turn induces an output voltage that changes in a continuous fashion. In this part, we consider a transistor as a controllable switch: the input signal is a low or high voltage, and the output voltage will be as well.
An npn transistor behaves as a closed switch when $v_{BE} > 0.6$ V. In that case there will be a lot of current i_C for a small v_{CE} and the transistor is in the saturation region. When $v_{BE} < 0.6$ V, there will be (almost) no current and the transistor is in the cut-off region. It acts as an open switch. Obviously, an pnp is open or closed when $v_{EB} < 0.6$ and $v_{EB} > 0.6$, respectively.
The situation is similar for a MOSFET: an n-channel MOSFET behaves as an closed switch when $v_{GS} > V_T$. The transistor will operate in the linear region - especially there where the curves are steep and the resistance is small. On the other hand, when $v_{GS} < V_T$, no current can flow and the MOSFET is an open switch. For a PMOS, the criterion is $v_{SG} > V_T$ (closed) or $v_{SG} < V_T$ (open).

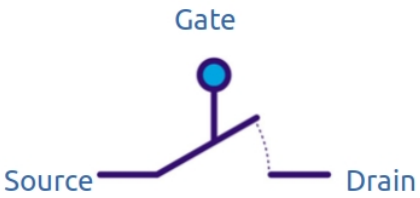


Figure 13.1

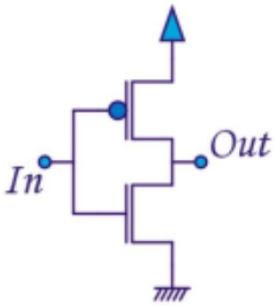


Figure 13.2

When we put a PMOS on top of an NMOS, we create a NOT-gate, as in figure 13.2. If the input is at a low voltage (a logical "0"), $V_{GS,N} < V_T$ and $V_{SG,P} > V_T$, so the bottom transistor is an open switch and the top one is closed. This means that the output is connected to the supply, and it is high ("1"). If the input is high, the top transistor will not conduct and is an open switch, and the output is connected to the ground via the closed NMOS. Hence the output is low. From this analysis, we can construct the truth table for this gate:

In	Out
0	1
1	0

Because the output is the opposite of the input, this is called a NOT-gate. The circuit is constructed such that both "switches" can never be closed at the same time. We will come back to this property in chapter 14.

We must however keep in mind that a logical gate is in essence an analog circuit, and the output v_{out} varies in a continuous way as the input voltage v_{in} increases, as in figure 13.3. We can analyze this behavior with the help of figure 13.4, where we see the $i_{DS} - v_{DS}$ characteristic the NMOS, and superimposed the $i_{SD} - v_{SD}$ curve of the PMOS. We can do this because $i_{DS} = i_{SD} = i$ and $v_{DS} + v_{SD} = E$.

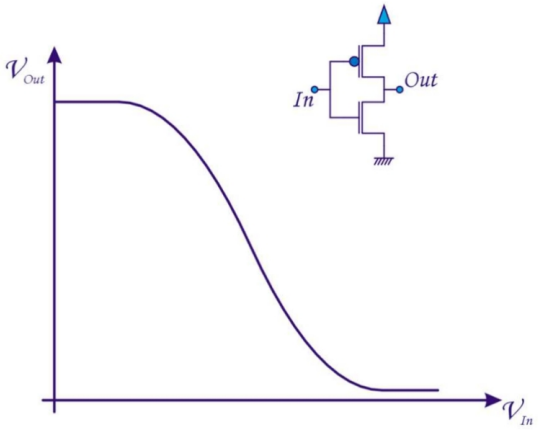


Figure 13.3

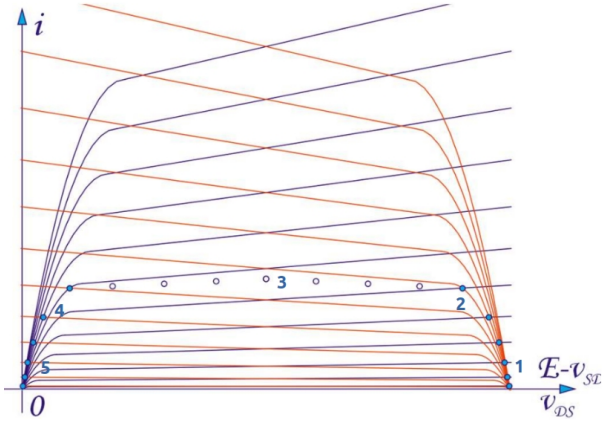


Figure 13.4

If v_{IN} is small, the NMOS is in cut-off , while the PMOS is in deep triode (linear) region because it wants to draw lots of current but it can't. We are in point ①, and the output voltage $v_{out} = v_{DS}$ is high. As v_{IN} increase and becomes $v_{IN} = v_{T,N}$, the NMOS will turn on and, because v_{DS} is high, will be in saturation. So we will be on a higher blue line which represents $i_{DS} = f(v_{DS})$ for a fixed $v_{GS,N}$. At the same time, $v_{SG,P}$ decreases so we will be on lower red line. The net result is that the intersection point (the dot) ¨moves slightly to the left and V_{out} starts to decrease slightly ②. At a certain moment, the PMOS leaves the linear region because $v_{SD} = E - v_{out}$ becomes large enough (i.e. $v_{SD} > v_{SG,P} - V_{T,P}$). From that moment on, there will be a significant current in the gate, and the output will decrease fast with increasing v_{IN}. We are in point ③ of figure 14.1. This continues until the NMOS goes in the linear region ④, where the current will decrease, and finally the input voltage becomes too high to create a channel in the PMOS: $v_{SG,P} < V_{T,P}$: there is no more current and the output is low ⑤.

This also means also that the input-output characteristic of the gate will depend on the I-V characteristics of the transistor, and consequently there will be a lot of variance between these curves. To accommodate this, the manufacturer will define certain threshold voltages to ensure proper operation of the gate, as in figure 13.5:

- V_{IH} : Minimum input voltage so that the output is guaranteed to be low,
- V_{IL} : Maximum input voltage so that the output is guaranteed to be high,
- V_{OH} : Minimum gate output when the gate generates a logic high
- V_{OL} : Maximum gate output when the gate generates a logic low

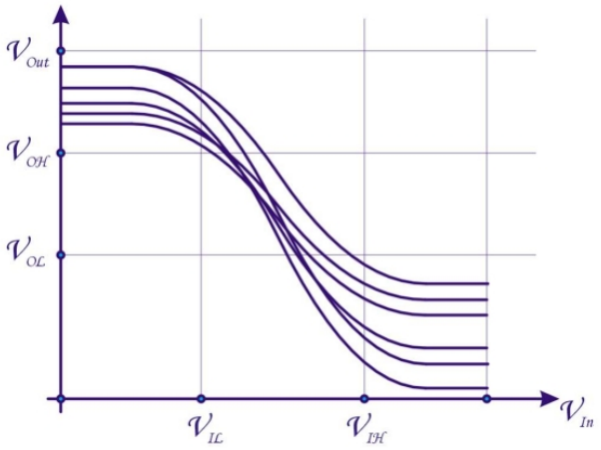


Figure 13.5

A specific technology will work only if $V_{IL} > V_{OL}$ and $V_{IH} < V_{OH}$, because only then can we put multiple gates in cascade and still guarantee their proper operation.
There exist different logic families:

- DTL (Diode-Transistor Logic)
- TTL (Transistor-Transistor Logic)
- ECL (Emitter-Coupled Logic)
- CMOS (Complementary MOS)
- ...

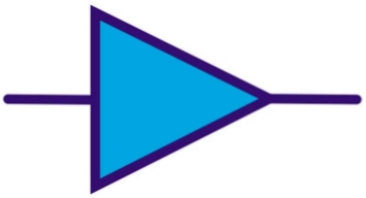
We will mostly work with CMOS where both NMOS and PMOS devices are used in the same substrate, but we'll also study other technologies in chapter 15. Each family has its own nominal power supply and threshold voltages V_{IH}, V_{IL}, V_{OH} and V_{OL}. It is therefore not recommended to mix different logical families. If this is required, you must check that $V_{IL} > V_{OL}$ and $V_{IH} < V_{OH}$, or use adapter gates.

Chapter 14

Logic Gates

14.1 Basic Logic Gates

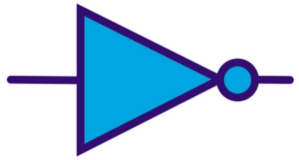
We already saw the NOT-gate in the previous chapter, together with its truth table. From boolean logic, we also know that we can create gates with more than one input, like the well-known AND- and OR-gates. These gates, together with their symbol are shown on the next page. We can also combine a NOT-gate with an AND- or OR-gate, to obtain a NAND and a NOR-gate. This may seem cursory, but in actuality it is easier to produce a NAND or NOR-gate in CMOS than a simple AND- or OR-gate. Furthermore, it can be shown that these gates are universal: any combinatorial circuit can be realized with only NAND or NOR-gates.
Another specific gate is the eXclusive OR or XOR-gate: its output is only high when exactly one of its inputs is high, and zero otherwise. This operation is also represented by the symbol $a \oplus b$.
Finally, a gate that produces the same logical output as its input is a buffergate, as shown in figure 14.1 together with its truth table, which is trivial. This gate must be used to interface with analog circuits, i.e. every time you receive a binary input somewhere, or when a binary output has to drive a (large) load. They should also be used when going from one logic family to another, like when you use TTL to drive a CMOS camera.



In	Out
0	0
1	1

Figure 14.1

In this chapter, we will use these gates to construct arbitrary boolean expressions. We'll see what Karnaugh mapping is and how it can be used to simplify expressions, and how these logic functions can be implemented with CMOS gates. We'll see two examples of this design procedure: the implementation of a digital multiplexer or MUX, and the implementation of an adder circuit.



NOT a $= \overline{a}$

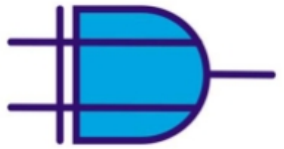
In	Out
0	1
1	0

a AND b $= a \cdot b$

a	b	$a \cdot b$
0	0	0
0	1	0
1	0	0
1	1	1

a OR b $= a + b$

a	b	$a + b$
0	0	0
0	1	1
1	0	1
1	1	1



a XOR b $= a\overline{b} + \overline{a}b$

a	b	$a\overline{b} + \overline{a}b$
0	0	0
0	1	1
1	0	1
1	1	0

a NAND b $= \overline{a \cdot b}$

a	b	$\overline{a \cdot b}$
0	0	1
0	1	1
1	0	1
1	1	0

a NOR b $= \overline{a + b}$

a	b	$\overline{a + b}$
0	0	1
0	1	0
1	0	0
1	1	0

Figure 14.2

14.2 Complex Logic Gates

Any logic function F can be defined with its truth table, where every combination of the inputs is enumerated together with the output. For a system with N inputs, this table has 2^N input lines[1]. As an example, consider the table in figure 14.3. Equivalently, we can represent a logic function by a Boolean expression. This can be done in two ways:

1. **Sum-of-Products**: only consider the rows with output equal to 1
   $F = \overline{A} \cdot \overline{B} \cdot C + \overline{A} \cdot B \cdot C + A \cdot \overline{B} \cdot \overline{C} + \overline{A} \cdot B \cdot C$

2. **Product-of-Sums**: only consider the rows with output equal to 0
   $F = (A + B + C) \cdot (A + \overline{B} + C) \cdot (\overline{A} + B + \overline{C}) \cdot (\overline{A} + \overline{B} + \overline{C})$

[1]The total number of possible truth tables for systems with N inputs is 2^{2^N}

A	B	C	F
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	0
1	1	0	1
1	1	1	0

Figure 14.3

These methods are equivalent:

$$\begin{aligned} F = \overline{\overline{F}} &= \overline{\overline{A}\cdot\overline{B}\cdot\overline{C} + \overline{A}\cdot B\cdot\overline{C} + A\cdot\overline{B}\cdot C + A\cdot B\cdot C} \\ &= \overline{\overline{A}\cdot\overline{B}\cdot\overline{C}}\cdot\overline{\overline{A}\cdot B\cdot\overline{C}}\cdot\overline{A\cdot\overline{B}\cdot C}\cdot\overline{A\cdot B\cdot C} \\ &= (A+B+C)\cdot(A+\overline{B}+C)\cdot(\overline{A}+B+\overline{C})\cdot(\overline{A}+\overline{B}+\overline{C}) \end{aligned}$$

where we applied the sum-of-products in the first line and the laws of de Morgan in the last two lines:

$$\begin{aligned} \overline{A\cdot B} &= \overline{A}+\overline{B} \\ \overline{A+B} &= \overline{A}\cdot\overline{B} \end{aligned}$$

to obtain the product-of-sums expression.
It is immediately clear that the sum-of-products expression can be implemented as in figure 14.4, while the product-of-sums is implemented in figure 14.5.

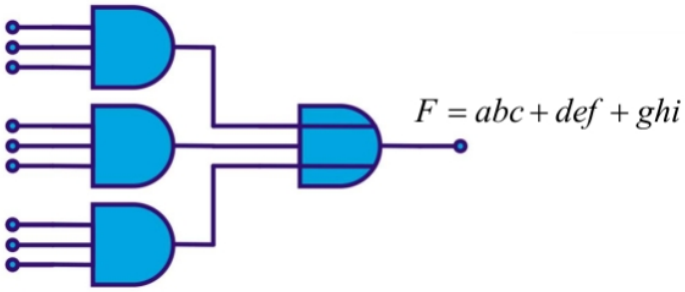


Figure 14.4

This is enough for implementation, but these circuits can be simplified by using only one type of logic gate. Any sum-of-products can also be implemented with only NAND-gates, as in figure 14.6, and a product-of-sums expression with only OR-gates.

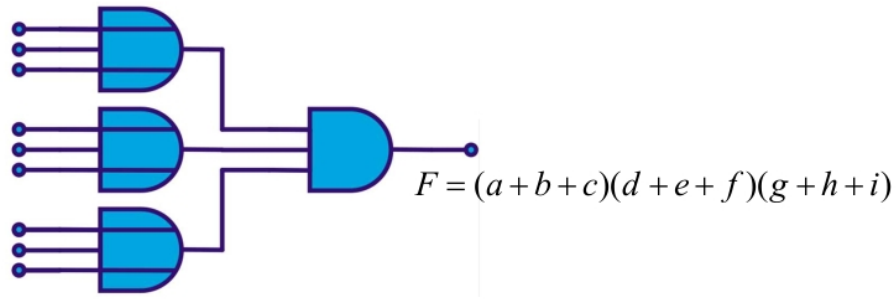


Figure 14.5

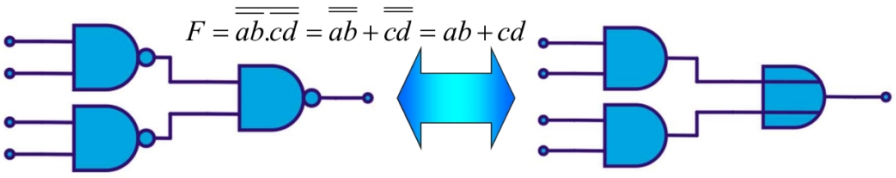


Figure 14.6

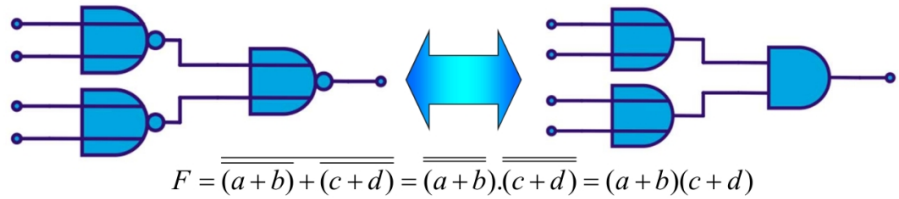


Figure 14.7

14.2.1 Example: Digital Multiplexer

We apply this method to the design of a digital multiplexer. A digital multiplexer or *MUX* is a device with 2 inputs A and B and a selector M. The single output of the MUX is equal to A if $M = 1$ and equal to B if $M = 0$. Its symbol is shown in figure 14.8. The truth table can be trivially constructed from the description and is represented in figure 14.9.

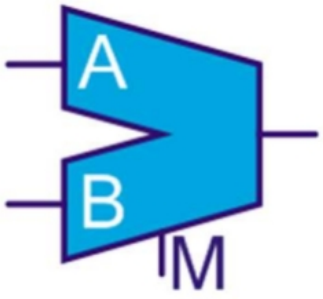


Figure 14.8

A	B	M	F
0	0	0	0
0	0	1	0
0	1	0	1
0	1	1	0
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

Figure 14.9

Based on the truth table, we can construct the sum-of-products:

$$\begin{aligned} F &= \overline{A} \cdot B \cdot \overline{M} + A \cdot \overline{B} \cdot M + A \cdot B \cdot \overline{M} + A \cdot B \cdot M \\ &= \overline{A} \cdot B \cdot \overline{M} + A \cdot B \cdot \overline{M} + A \cdot \overline{B} \cdot M + A \cdot B \cdot M \\ &= (\overline{A} + A) \cdot B \cdot \overline{M} + (\overline{B} + B) \cdot A \cdot M \\ &= B \cdot \overline{M} + \cdot A \cdot M \end{aligned}$$

because of the distributive property and $\overline{X} + X = 1$ for any X. It should be clear that simplifying these equations is important to reduce circuit complexity. We will see how we can do this reduction systematically in section 14.4.
To implement the logic function, we need two AND-gates, one OR-gate and a NOT-gate to generate $\overline{M}$. The result is shown in figure 14.10.

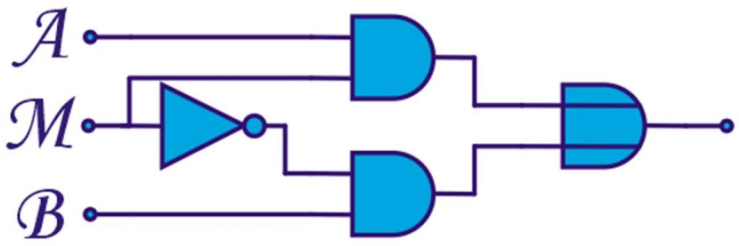


Figure 14.10

14.3 CMOS Gates

We already saw how we can make a NOT-gate from an NMOS and a PMOS transistor - see figure 13.2. To make a NAND- or NOR-gate, we need additional transistors to accept two input signals. How this is done in shown in figure 14.11 (NAND) and 14.12 (NOR).

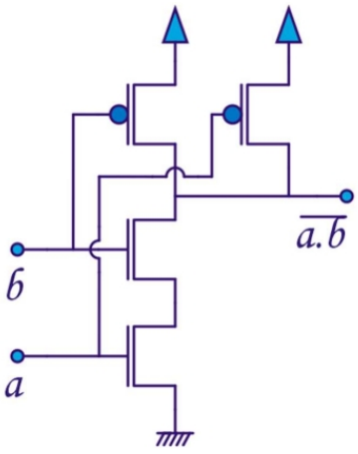


Figure 14.11: CMOS NAND Gate

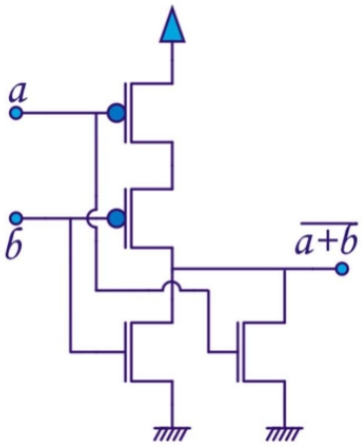


Figure 14.12: CMOS NOR Gate

For the NAND-gate, the output is pulled to ground only when both inputs are high because then both NMOS in series will conduct. If at least one input is low, a PMOS in the top part will conduct and pull the output to the supply (high).
For the NOR-gate, the situation is reversed: if one input is high, the output is pulled to ground. If both are low, both PMOS in series conduct and the output is high.
The extension to more terminals is trivial, because it only requires the addition of more transistors, either in parallel or series with the existing transistors.
Note how these circuits have the same general structure: there are always two distinct blocks: the block on the top (with the PMOS transistors) is connected to the supply. When this block in total is closed the output is "1". If not, it is an open switch (high impedance). For the bottom part, containing the NMOS transistors, the situation is reversed: this part of the circuit generates the logic "0" because it can connect the output to ground. It is only a closed switch when the output is not "1", and an open switch otherwise. This means that both blocks have to be complementary. If this would not be the case, certain input configurations could create a conducting path between supply and ground.

This duality is represented in figure 14.13, where the top block corresponds to $F = 1$, and

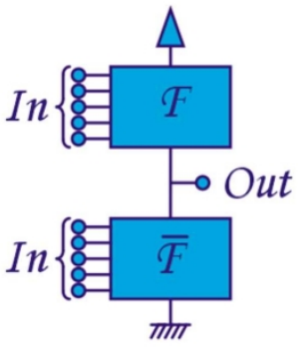


Figure 14.13

the bottom block to $\overline{F} = 1 \Leftrightarrow F = 0$. For the NAND-gate:

- p-block: $\overline{a} + \overline{b} = \overline{a \cdot b}$
- n-block: $a \cdot b$

a	b	$a\overline{b} + \overline{a}b$
0	0	0
0	1	1
1	0	1
1	1	0

Table 14.1

And for the NOR-gate:

- p-block: $\overline{a} \cdot \overline{b} = \overline{a + b}$
- n-block: $a + b$

This principle can be used to make gates of arbitrarily complexity. For instance, consider the XOR-gate we've seen before and which has the truth table in table 14.1. The expression for the logic function is $F = \overline{a} \cdot b + a \cdot \overline{b}$.

This expression for F determines the p-block of the circuit, consisting of two branches in parallel, where each branch has two PMOS in series: $\overline{a} \cdot b$ and $a \cdot \overline{b}$.
To determine the n-block, we compute $\overline{F}$ by looking at the zero outputs: $\overline{F} = \overline{a} \cdot \overline{b} + a \cdot b$. This expression also leads to two parallel branches, each containing two NMOS in series, between output and ground. Both p- and n-blocks are implemented in figure 14.14.

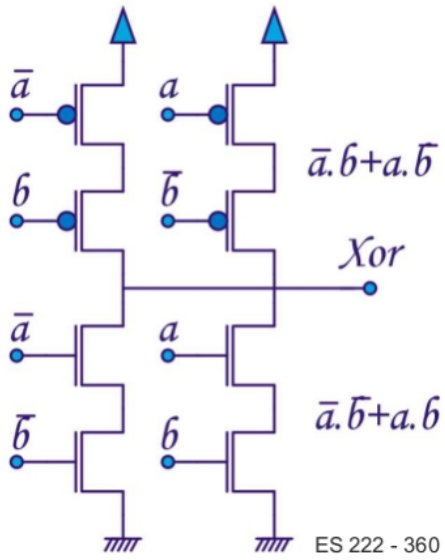


Figure 14.14

14.3.1 Gate Inhibition

It can be a problem when multiple gates are connected to the same point. For instance, consider the bus topology in figure 14.15, where multiple NAND-gates are connected to the same bus. If the output of one gate is high and the other gate is low, we have created a short-circuit between supply and ground. To avoid this, we add an extra input pin that allows to

inhibit or deactivate the gate. Concretely, it will open the connections to both supply and ground, as in figure 14.16 so the output is left dangling and its value is undetermined.

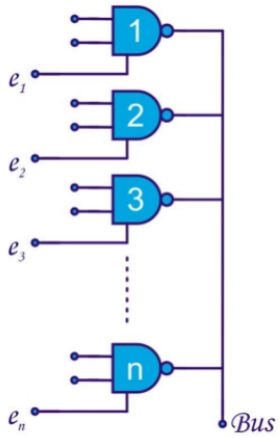


Figure 14.15

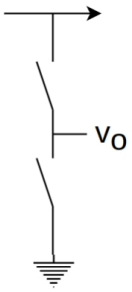


Figure 14.16

To implement this in CMOS, we put additional P- and NMOS-transistors between the output and respectively the upper and lower part, as in figure 14.17. The PMOS is driven by the inhibitor signal e and the NMOS by $\overline{e}$. So if e is high, both upper and lower parts are isolated from the output. This is *active high* inhibition.

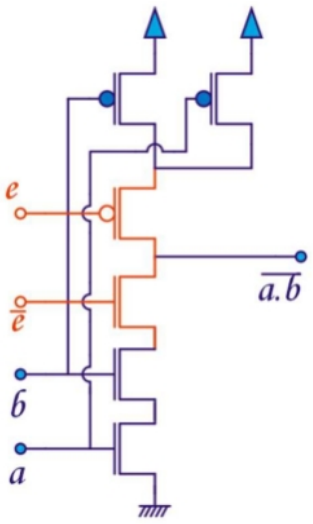


Figure 14.17

14.4 Karnaugh Mapping

Any logic function F can be uniquely defined with a truth table. However, the number of entries in this table depends exponentially on the number of inputs, as we've seen before. It is therefore imperative to not only obtain a logic expression from this table, but also to reduce this expression as much as possible before implementation. This reduces the number of components, and hence the complexity and chance of failure of the circuit.

A consistent way to reduce a boolean expression to a simpler form was proposed by Maurice Karnaugh in 1953. As an example, we consider a function of two variables:

$$F = a\overline{b} + ab + \overline{a}b$$

This function can be represented in a Karnaugh table, which is similar to the truth table we used before, but with the input variables rearranged in rows and columns. In figure 14.18, input a is shown in the columns, input b in the rows. The trick in the Karnaugh table is to group all outputs of the same kind ("0" or "1") in groups of size 2^k, until all these similar outputs belong to at least one group. The goal is to choose these groups as large as possible. In the figure, we need 2 groups of size 2^1 to collect all the "1"s. Karnaugh shows that the expression can be reduced to $F = a + b$, because we find that $F = 1$ when either $a = 1$ (group on the right) or if $b = 1$ (group on the bottom).

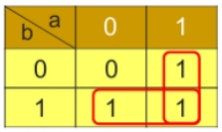


Figure 14.18

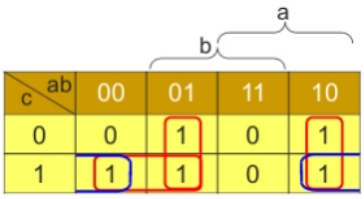


Figure 14.19

Could this simplification be achieved by symbolic manipulation? We could have grouped different terms in our expression:

$$F = a\overline{b} + ab + \overline{a}b = a(\overline{b} + b) + \overline{a}b = a + \overline{a}b$$

but we can go no further in this way. But with the Karnaugh map, we see that we need to reuse a term (ab) to improve this result:

$$F = a\overline{b} + ab + ab + \overline{a}b = a(\overline{b} + b) + (\overline{a} + a)b = a + b$$

In summary, the Karnaugh map shows us what terms to reuse, and with which other terms they should be recombined.

When we have three inputs, we group multiple inputs in a column, as in figure 14.19. It is important to write these inputs so that only a single bit changes from column to column - this is called a *Grey code*. Once again, we group elements in groups of size 2^k - we find three different groups: $F = a\overline{b} + \overline{a}b + \overline{a}c$. Note that we could also have chosen another group: $F = a\overline{b} + \overline{a}b + \overline{b}c$.

If we have 4 input variables, we encode them also in the rows - once again with a Grey code - and apply the exact same method. This is the example in figure 14.20, where the grouping already has been done. The result is $F = \overline{ab} + \overline{ac} + \overline{ad} + bcd + \overline{bc} + \overline{bd}$.

With more than 4 variables, the visualization becomes harder. With 5 variables, you can use two tables: one for the last variable = 0, and the other for = 1, but remember you'll have to group across tables. For many variables, you should use specialized software tools.

14.4.1 Example: the Adder

We now build a digital circuit that adds two binary numbers a and b. Furthermore, it can also accept a "carry in" c_{in} so it can be used in a cascade of similar circuits that will form

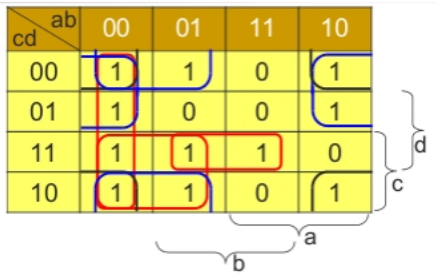


Figure 14.20

an adder. It has two outputs: the sum s and the "carry out" c_{out}, as in figure 14.21. The sum is $s = a \oplus b \oplus c_{in}$ with $\oplus$ the sum modulo 2, and $c_{out} = \lfloor (a + b + c_{in})/2 \rfloor$. The results are summarized in the truth table in figure 14.22. Since we have two outputs, we need to construct two Karnaugh tables: one for c_{out}, as in figure 14.23, and one for s in figure 14.24. The table for c_{out} gives:

$$c_{out} = a \cdot c_{in} + b \cdot c_{in} + a \cdot b$$

In the table for s, we find only groupings of size 1, so s can't be simplified beyond

$$s = a \oplus b \oplus c_{in}$$

.

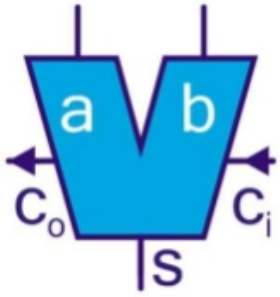


Figure 14.21

a	b	c_{in}	s	c_{out}
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

Figure 14.22

c_{in} \ ab	00	01	11	10
0	0	0	1	0
1	0	1	1	1

Figure 14.23: Table for c_{out}

c_{in} \ ab	00	01	11	10
0	0	1	0	1
1	1	0	1	0

Figure 14.24: Table for s

We implement s with two XOR-gates, as in figure 14.25. Output c_{out} can be constructed with 4 NAND-gates (confirm this) as in figure 14.26.

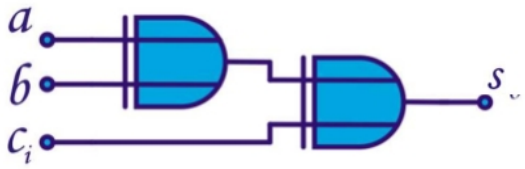


Figure 14.25

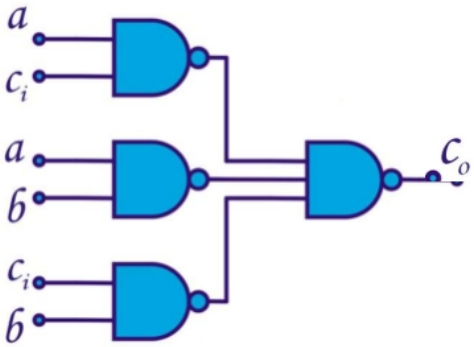


Figure 14.26

When we put multiple of these circuits in cascade, we get an adder. For example, to add 2 4-digit binary numbers, we can use the circuit in figure 14.27 where the c_o of one element is the c_i of the next one. The carry in for the initial lowest bit is zero. If the sum can not be represented by a 4-bit number, there will be an *overflow* bit.
The circuit can also be used to subtract, because $b-a = b+(-a)$ and we can easily construct $-a$ with the two's-complement: for any binary $a = \sum_{i=0}^{n-1} b_i 2^i$:

$$-a = 2^n - \sum_{i=0}^{n-1} b_i 2^i = (1 + \sum_{i=0}^{n-1} 2^i) - \sum_{i=0}^{n-1} b_i 2^i = 1 + \sum_{i=0}^{n-1} (1 - b_i) 2^i$$

So we construct $-a$ by inverting each bit and adding 1, as in figure 14.28, where the 1 is added through the initial carry in.

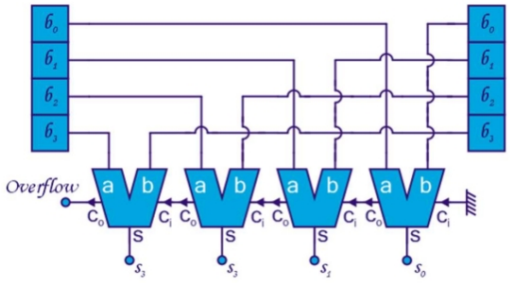


Figure 14.27

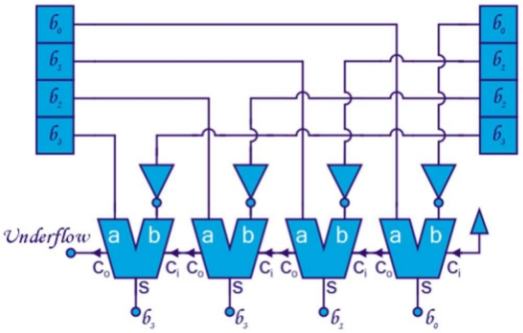


Figure 14.28

Chapter 15

Alternative Digital Families

In chapter 14, we used Complementary MOS to implement the digital circuits. This is the predominant technology in use today. However, there are some applications that require specific digital technologies, mostly based on bipolar transistors. These technologies are much faster than CMOS.
In this chapter, we will discuss the most important of these BJT digital technologies: *Transistor-Transistor Logic* or TTL, and ECL or *Emitter-Coupled Logic*. At the end of the chapter, we'll also have a look at a special purpose circuit: the Schmidt trigger.

15.1 BJT Logic Gate

The first logic gates where implemented with bipolar transistors. A basic BJT circuit to implement a NOT-gate is seen in figure 15.1. A high input voltage will generate a high base current through resistor R_B and as a consequence, there will be a high collector current through R. The output $E - R\ I_C$ will be low. If the input is low, there will be no base or collector current and the output is high. This technology is also called *Resistor-Transistor Logic* (RTL).

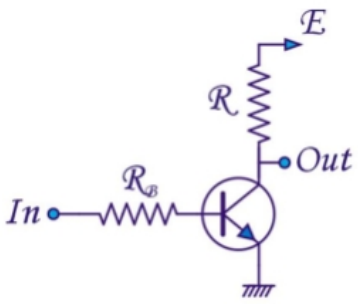


Figure 15.1

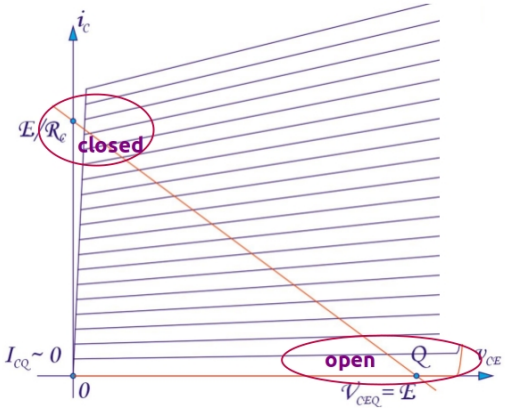


Figure 15.2

When the input is high ($\approx E$), there is lots of current and the transistor is in saturation and behaves as an closed switch. When the input is low (≈ 0), there is no base current, the transistor is in cut-off and behaves as an open switch - see figure 15.2. The input-output

characteristic is seen in figure 15.3 and is as expected.

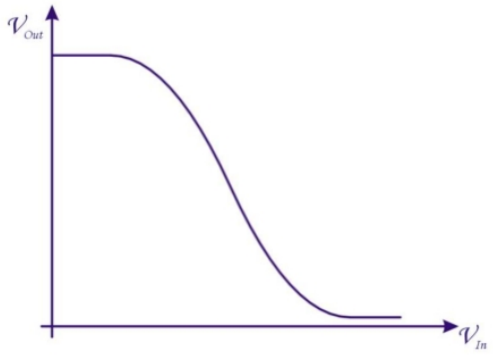


Figure 15.3

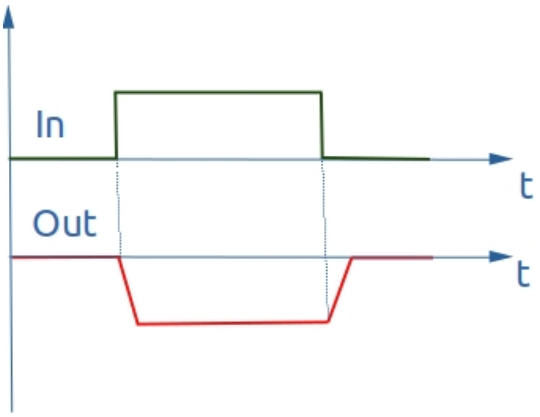


Figure 15.4

To analyze the dynamic behavior of the circuit, we need to take the capacitances at in- and output into account. This means that the output transient behavior doesn't follow the input immediately, but its speed is determined by the charging and discharging of the (parasitic) capacitance at rate $\frac{I}{C}t$ (for a constant current), as in figure 15.4 with the slope of the output flanks $= \frac{I}{C}$. We need lots of current for a very fast gate.
The input capacitance is always present due to the junction capacitance C between base and emitter. The current we can inject into the base is $I_B = \frac{E-V_{BEQ}}{R_E}$. This is the current available to charge C and get a low output. To discharge, we need to put the input at ground, and the current is $I_B = \frac{V_{BEQ}}{R_E}$ to get a high output. This current is a lot lower than the charge current, what means that the gate goes very fast from 1 to 0, but very slowly from 0 to 1. To improve this behavior, we replace the resistor by a transistor.

15.2 Transistor-Transistor Logic

By replacing the resistor by a transistor, we work with *Transistor-Transistor Logic* (TTL). If the input in figure 15.5 is low, the base-emitter junction of transistor ② is forward biased because it sees a voltage E, and you will draw a current through the base via R_B:

$$I_B = \frac{E - V_{BEQ}}{R_B} \text{ and } I_C = \beta\frac{E - V_{BEQ}}{R_B}$$

at least until the base capacitance is discharged. This current is lot higher than $\frac{V_{BEQ}}{R_E}$ we had in RTL. Transistor ① from which we try to draw current from the base will block and the output is high.
When the input is high, we are in the peculiar situation where we are using the input transistor in inverted mode (see section 5.1.4) where emitter and collector are interchanged. This means that between supply and ground we have two V_{BEQ}'s: once for the input transistor ② via R_B and with the collector-base junction in forward bias, and between base and emitter of transistor ①. The base current of transistor ② is $\frac{E-2V_{BEQ}}{R}$. While a transistor doesn't work as well in inverted mode, there is still current amplification (but with reduced β) so there is more than enough current available to charge the parasitic capacitance of transistor ① and to polarize ① in the saturation domain, leading to a low output.

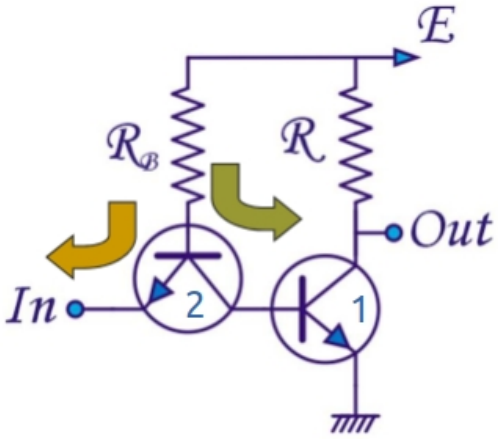


Figure 15.5

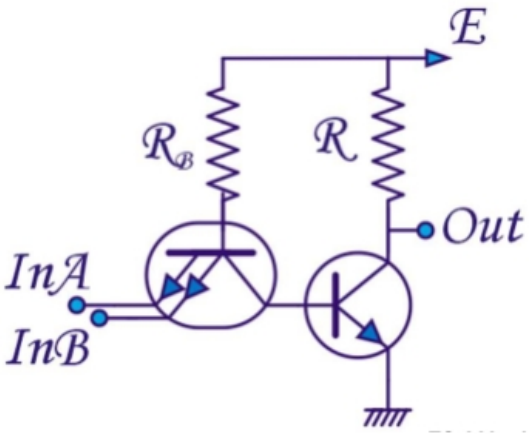


Figure 15.6

To create a NAND-gate, we replace transistor ② with 2 BJT transistors in parallel (i.e. base and collector at the same voltage) - as drawn in figure 15.6. The two emitters of these parallel input transistors will be the inputs. If one of the two inputs is low, its transistor will try to draw a current out of the base of the output transistor, such that this one will be in cut-off and the output is high. If both inputs are high, there is a large base current into the output transistor, so the output is low, as required.

15.2.1 Totem-Pole Configuration

The input transistor ② in 15.3 solved the issue with time difference between charging and discharging due to the input capacitance. We still have to deal with the output capacitance C_L due to the load, as in figure 15.7. This capacitor charges through resistor R - with a time constant $R\,C_L$ - and discharges through the BJT. The issue is that discharging via the BJT can happen a lot faster than charging via R.

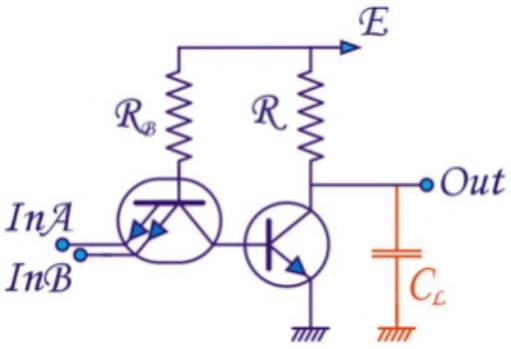


Figure 15.7

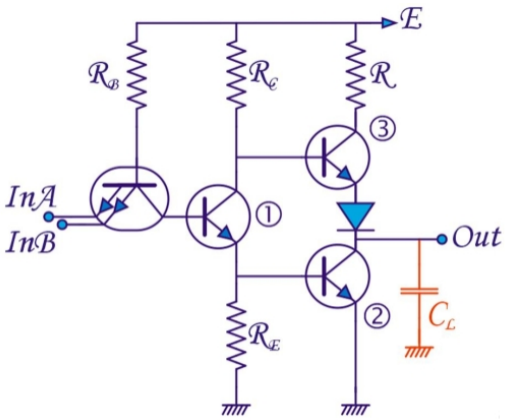


Figure 15.8

We improve this by:

- Modifying the output stage by adding an emitter resistance R_E. This is now a phase-splitter.
- Adding an output stage with 2 BJT on top of each other. This is a *totem-pole* configuration. The load capacitance C_L will now charge through transistor ③ and discharge

through transistor ②.

With ① operating as a phase splitter, we can relate the voltage variations at the bases of transistors ② and ③:

$$\frac{v_{be3}}{v_{be2}} = -\frac{R_C}{R_E} \Leftrightarrow \frac{v_{be2}}{R_E} = -\frac{v_{be3}}{R_C}$$

If an input is low, it tries to draw current from transistor ① and v_{B1} goes down, as in figure 15.9. Transistor ① is an open switch and we can remove him for the analysis. because transistor ② has $V_{BEQ} = 0$, it will block and ③ will saturate and conduct, so the output is high and C_L charges through transistor ③. However, the output is not equal to the supply E - it is set via R_C and the voltage drops over the base-emitter junction of ③ and the diode: $v_{OUT} = E - R_C I_B - V_{BEQ3} - V_{DQ} \approx E - 1.2V - R_C I_B$. Resistance R_C can not be too high to reduce this voltage drop. For $E = 5$ V, the output will be ≈ 3.5 V.

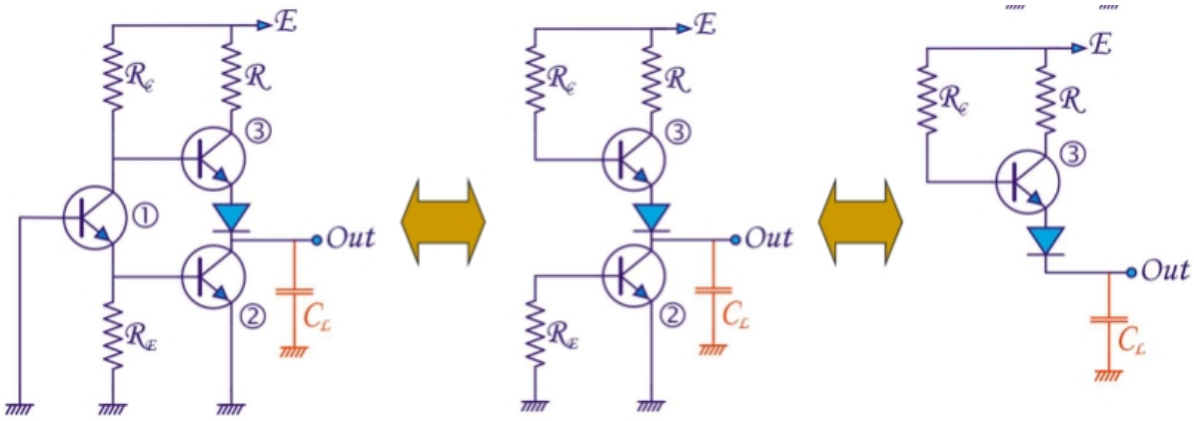


Figure 15.9

When both inputs are high, we can replace their base-collector junction by a forward-biased diode, as in figure 15.10. Transistor ① is saturated and we can replace it by a voltage source $V_{CE,Sat}$. This means that $V_{B3} = V_{BEQ2} + V_{CE,Sat1} \approx 0.8$ V. Resistance R_E is chosen such that $V_{BE2} = (E - V_{CE,Sat})\frac{R_E}{R_E+R_C} \approx 0.8$ V. Consequently $V_{BE3} = V_{B3} - V_{E3} = V_{BEQ2} + V_{CE,Sat1} - (V_{CE,Sat2+V_{DQ}}) \approx 0$. So transistor ② is saturated and ③ is blocked. This is also why the diode is there: to make sure that ③ is blocked when ② is saturated. The minimum output voltage that corresponds to a low signal is $V_{CE,Sat2} \approx 0.2$ V.

When a transistor turns on, we provide a lot of base current in TTL, more base current than it needs for the collector current it is drawing - the transistor is saturated. The extra base current creates a stored charge in the base of the transistor. As discussed, this stored charge causes problems when the transistor needs to be switched from on to off: while the charge is present, the transistor is on; all the charge must be removed before the transistor will turn off. Removing the charge takes time, so the result of saturation is a delay between the applied turn-off input at the base and the voltage swing at the collector. To avoid that the transistor saturates, we place a Schottky diode with a threshold voltage of 0.3 V between base an collector (figure 15.11). By doing so, the collector-emitter voltage $V_{CE} = V_{BEQ} - 0.3V = 0.3V > V_{CE,Sat}$ and the transistor can never saturate.

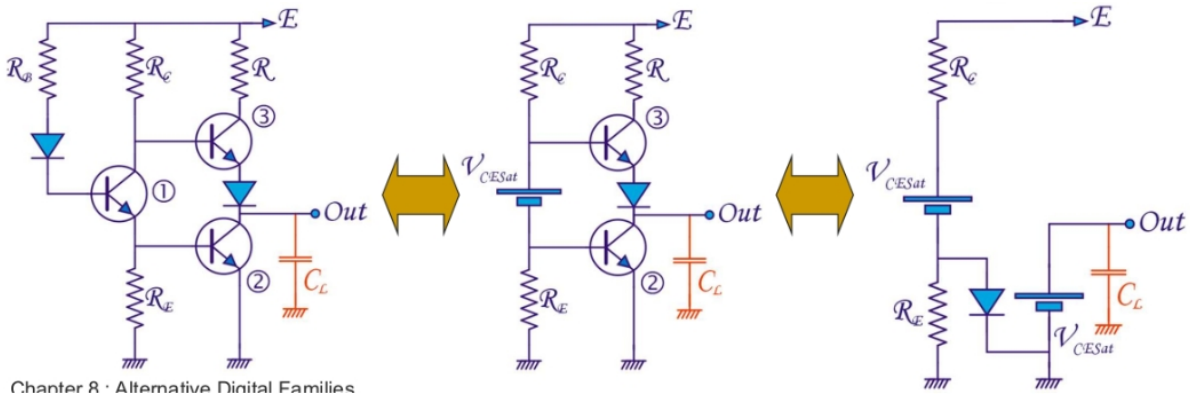


Figure 15.10

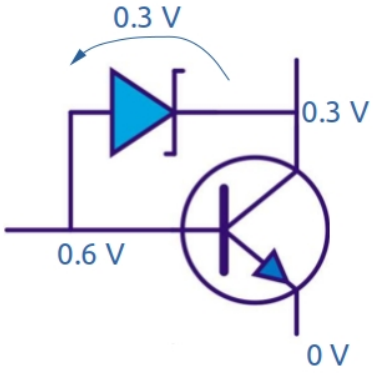


Figure 15.11: A Schottky transistor

15.3 Emitter-Coupled Logic

Transistor-Transistor logic is based on the saturation and blocking of the bipolar transistor (figure 15.2). This means that there will be large voltage swings and several crucial circuit node have to be charged and discharged. That's why TTL is inherently slow (but still faster than CMOS). *Emitter-Coupled Logic* (ECL) is faster because the transistors never saturate and the voltage swings are kept low ($\sim \pm 400mV$). The technology is also called *Current Mode Logic* (CML). It is based on a differential topology, as in the differential amplifier from chapter 7.3 where two input transistors share the emitter node.
A typical ECL NOT-gate is shown in figure 15.12. The true gate is the differential amplifier in blue. The two common-collector amplifiers in red are just there shift the outputs one V_{BEQ} lower and to buffer the outputs voltages because they provide a low output impedance. Note that one input transistor is put at a reference voltage V_{ref}. The other transistor provides the input.
When $V_a > V_{ref}$, all current I_{CC} will flow through the left branch. So the collector voltage $V_{C1} = E - R\ I_{CC}$ will be low and $V_{C2} = E$ will be high. We will assume that the input voltage is high when $V_a = V_{ref} + \frac{1}{2}V_{BEQ}$ and low when $V_a = V_{ref} - \frac{1}{2}V_{BEQ}$. At the output, we then have:

- HIGH $= E - V_{BEQ}$

- LOW $= E - R\, I_{CC} - V_{BEQ}$

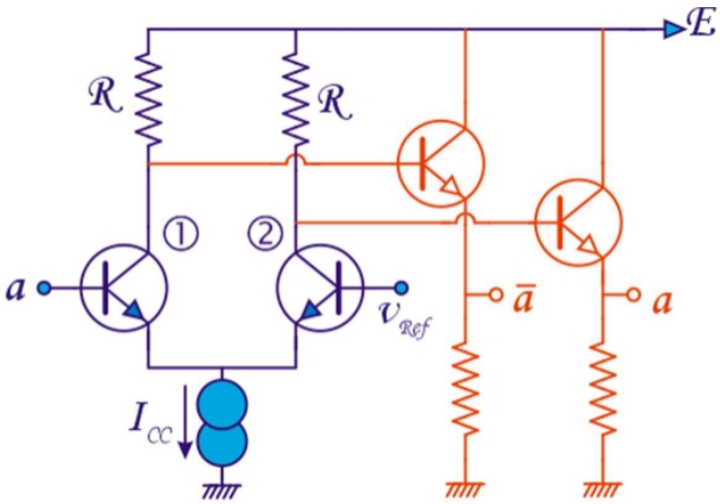


Figure 15.12

The difference between high and low (the output swing) is thus $R\, I_{CC}$ and must be equal to V_{BEQ} because this is the swing at the input. So, this means that:

- HIGH $= E - V_{BEQ}$
- LOW $= E - 2V_{BEQ}$

And the reference value must be halfway between HIGH and LOW: $V_{ref} = \frac{V_{HIGH}+V_{LOW}}{2} = E - \frac{3}{2}V_{BEQ}$. Note that the ECL-gate always has two complementary outputs. Furthermore, the transistor never saturates because $V_{CE,min} = V_{BEQ}$. This is because when $V_a = V_{ref} + \frac{1}{2}V_{BEQ}$, $V_{C1} = E - V_{BEQ}$ and $V_{E1} = V_{ref} - \frac{1}{2}V_{BEQ} = E - 2V_{BEQ}$, so $V_{CE1} = V_{BEQ} > V_{CE,Sat}$. This is even more valid for the other transistor, because his collector voltage is E, so $V_{CE} = 2V_{BEQ} > V_{CE,Sat}$.

15.3.1 NOR Gate

When we want to create a NOR-gate, it is enough to add a transistor in parallel with the existing input transistor, as in figure 15.13. If one of the inputs is HIGH, that branch will conduct (regardless of what happens with the other input) and there will be a voltage drop across R such that the voltage in the left node will be LOW. The other branch automatically offers a OR $b = a + b$.

15.3.2 NAND Gate

To implement a NAND-gate, you would add a transistor in series with the existing input transistor. In that case, there's only conduction in the left branch - and also a low voltage - when both inputs are high. However, the issue is that both inputs are not on the same level because the transistors are stacked. So we compare $a - V_{BEQ}$ with $V_{ref} - V_{BEQ}$ instead of comparing a with V_{BEQ}. Sa a is level-shifted down with one V_{BEQ} with the transistor to which a is connected.

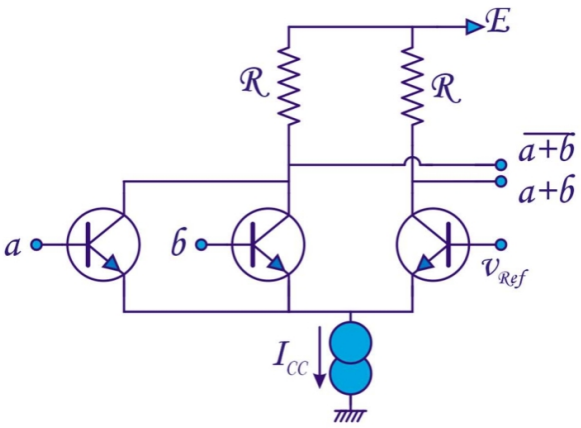


Figure 15.13

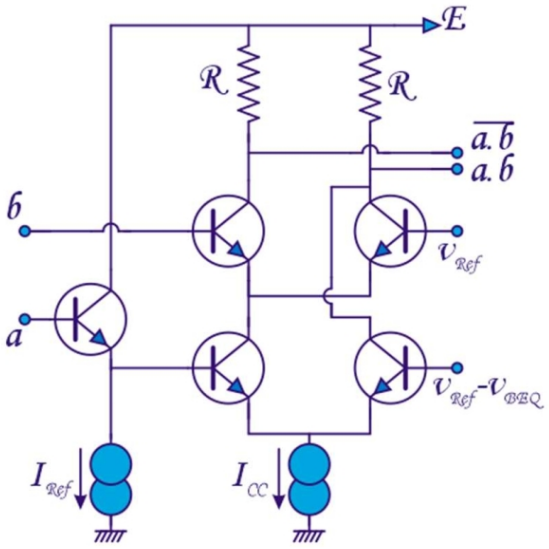


Figure 15.14

15.4 The Schmidt Trigger

A digital signal gets corrupted by noise, as in figure 15.15. Suppose you want to reconstruct the original signal. Conceptually, this should be simple: just use a comparator with a hard threshold (i.e. an OPAMP with no feedback and $v^- = v_{treshold}$) as in figure 15.16 where $v_{threshold}$ is midway between a LOW and HIGH voltage, and you should be able to remove the noise. However, as figure 15.17 demonstrates, this is not the case: the amplitude noise is transformed in phase noise, and the reconstructed signal is useless.

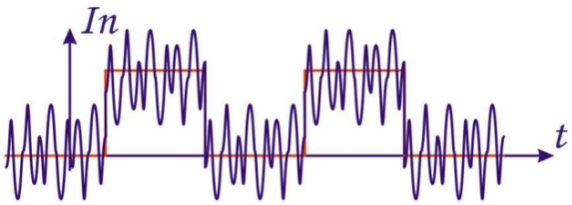


Figure 15.15

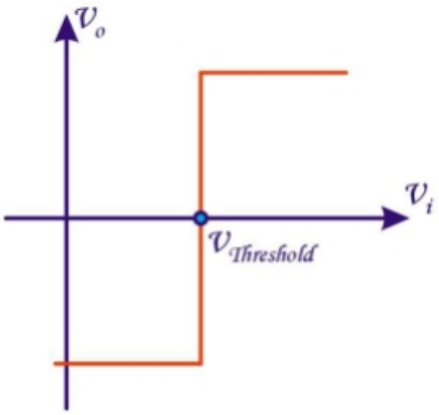


Figure 15.16

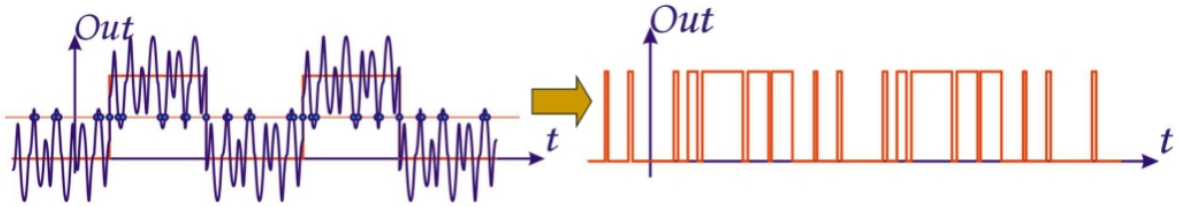


Figure 15.17

A better way to recover the original signal, is by using *hysteresis* - as shown in figure 15.18: a signal value is classified as HIGH when the voltage is higher than v_h. The signal will remain HIGH until it falls below a voltage $v_l < v_h$, when it will be considered as LOW. It will only return to being HIGH when $v_o > v_h$. When we apply this hysteresis comparator to the noisy signal, we can reconstruct the underlying signal perfectly, as figure 15.20 demonstrates.

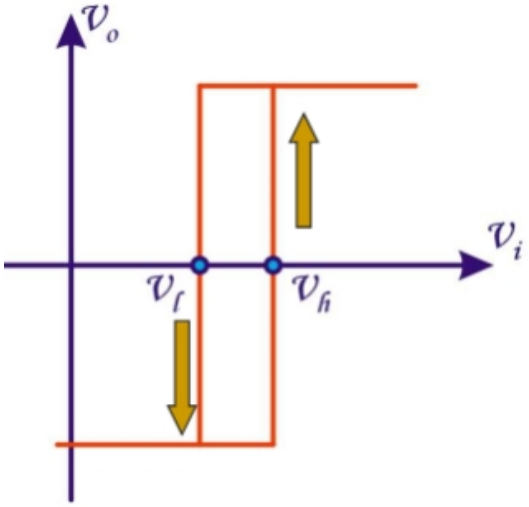


Figure 15.18

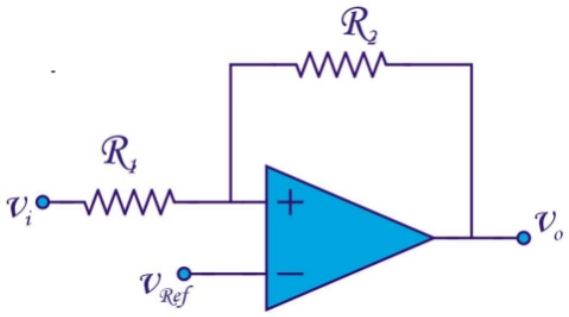


Figure 15.19

To implement a comparator with hysteresis, we use a Schmidt trigger: an OPAMP with feedback on the positive node, as in figure 15.19, so it is unstable. Note that we already used this device in the fantastron oscillator from chapter 11.6 but we will study it in more detail

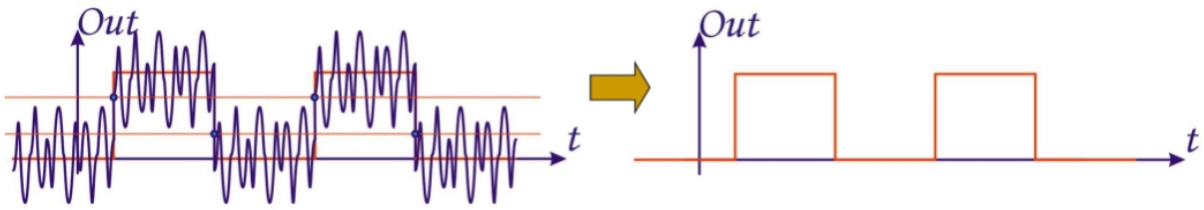


Figure 15.20

in this section.
At the positive node, we can write:

$$v_{ref} + v = \frac{v_i/R_1 + v_o/R_2}{1/R_1 + 1/R_2} \tag{15.1}$$

$$= \frac{\frac{R_2}{R_1} v_i + v_o}{1 + R_2/R_1} \tag{15.2}$$

$$= \frac{k v_i + v_o}{1 + k} \tag{15.3}$$

$$v_o = (1 + k)\, v_{ref} + (1 + k)\, v - k v_i \tag{15.4}$$

We can plot this line on the $v_o - v$ characteristic of the OPAMP: $v_o = \Phi(v)$, as in figure 15.21. The diagonal line that represents equation 15.4 moves horizontally with varying input voltage v_i. The intersection of this line with $\Phi(v)$ provides at maximum 3 solutions. The method of Lyapounov allows us to check the stability of these nodes by linearizing Φ around them: $v_o = A\, v$ with A very large in the middle and ≈ 0 when $v_o = \pm E$. If we work with a fixed v_i and fixed v_{ref}, we can also write 15.4 around the operating point:

$$v_o = (1 + k)\, v \tag{15.5}$$

With the OPAMP as a low-pass first-order system, the gain is

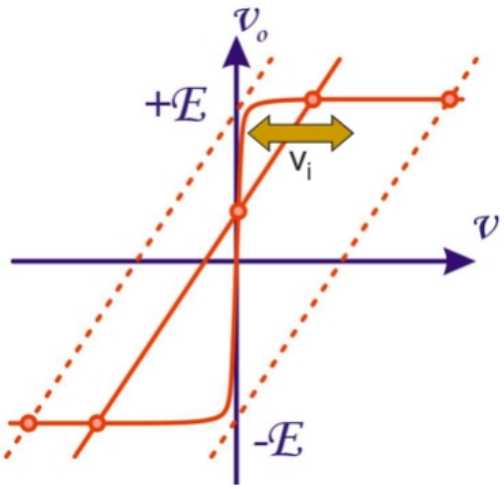


Figure 15.21

$$v_o = \frac{A_0}{1+sT}$$

Substituting this in equation 15.5, we get:

$$A_0 v_o = (1+k)v + sT(1+k)v = (1+k)v + T(1+k)\frac{d}{dt}v$$
$$\Rightarrow \frac{d}{dt}v = \frac{A_0 - (1+k)}{T(1+k)}v$$

So the circuit is stable when

$$A_0 - (1+k) < 0 \Leftrightarrow A_0 < 1+k$$

The system is stable in the operating points where $v_o = \pm E$ and unstable in the middle. This means that if the operating point ends up in the middle where A_0 is high, it will return to one of the two stable operating points after some time.
When v_i is high, the diagonal line lies to the right, and there is only one intersection with Φ, namely when $v_o = E$, and this point is stable. As v_i decreases, there will come a point when there is also an intersection with the lower part of Φ, namely when $v = 0$ with $v_o = E$, i.e.:

$$v = 0 \Rightarrow v_o = (1+k)v_{ref} - kv_i \Rightarrow v_i = \frac{R_1}{R_2}E + (1+\frac{R_1}{R2})v_{ref}$$

From that point one there are 3 intersections, but the operating point remains on $v_o = E$ because this point is stable. When $v_i = -\frac{R_1}{R_2}E + (1+\frac{R_1}{R2})v_{ref}$, the top operating point disappears because there is no longer an intersection with $v_o = E$ and the operating must move to the only remaining (stable) operating point , namely $v_o = -E$. When v_i starts to increase, the output remains equal to $-E$ as long as this point exists as a fixed point, only to switch to $v_o = E$ when it disappears. This cycle is represented in figure 15.22.

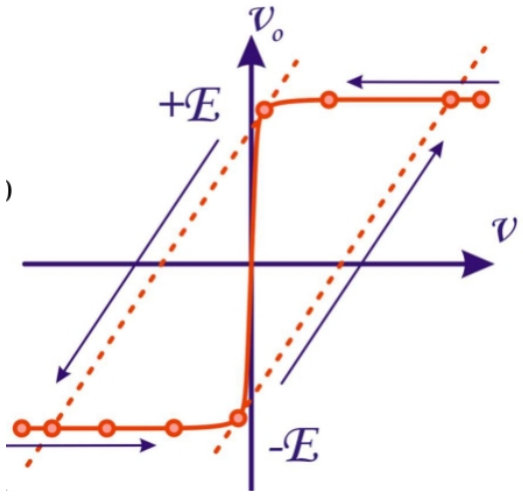


Figure 15.22

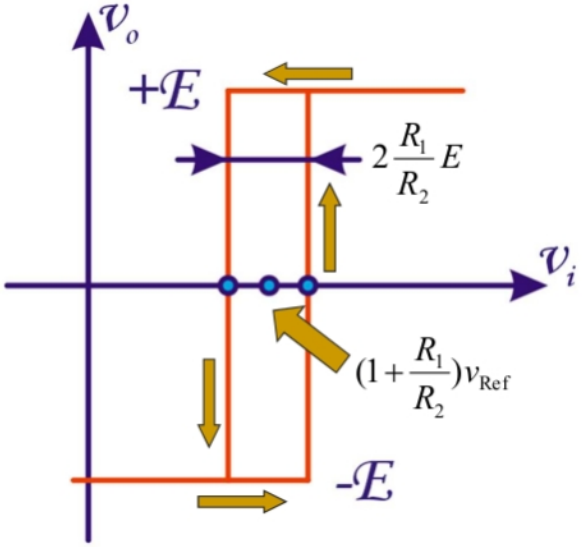


Figure 15.23

So we have created a hysteresis comparator with

$$v_l = -\frac{R_1}{R_2}E + \left(1+\frac{R_1}{R2}\right)v_{ref} \text{ and } v_h = \frac{R_1}{R_2}E + \left(1+\frac{R_1}{R2}\right)v_{ref}$$

as in figure 15.23. The trip point is $(1 + \frac{R_1}{R_2})\ v_{ref}$ and the hysteresis is $2\frac{R_1}{R_2}\ E$. We can use v_{ref} to set the trip point and $\frac{R_1}{R_2}$ to determine the hysteresis.
The symbol of this gate is represented in figure 15.24.

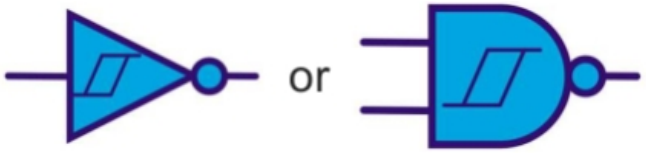


Figure 15.24

Chapter 16

Digital Circuit Implementation

16.1 Field-Programmable Gate Array

A field-programmable gate array (FPGA) is an integrated circuit that can be programmed and reprogrammed to perform a specific function or task. Unlike traditional application-specific integrated circuits (ASICs), which are designed for a specific purpose and cannot be modified after manufacturing, FPGAs are designed to be flexible and can be reconfigured or programmed by the user to perform a variety of tasks.

FPGAs consist of programmable logic blocks and interconnects that can be configured to create custom digital circuits. The programmable logic blocks can be configured to perform basic logic functions such as AND, OR, and XOR gates, and the interconnects allow these blocks to be connected to create more complex circuits.

FPGAs are commonly used in a variety of applications, including digital signal processing, image and video processing, networking, and aerospace and defense systems. They offer high performance, low power consumption, and flexibility, making them a popular choice for many types of applications.

16.2 Programmable Logic Array

A programmable logic array (PLA) is a type of digital logic device that is used to implement combinational logic circuits. It consists of an array of AND gates, followed by an array of OR gates. The inputs to the AND gates are programmable, meaning that they can be configured to take on any combination of values. The outputs of the AND gates are then fed into the OR gates, which produce the final output.

The configuration of the inputs to the AND gates is typically stored in non-volatile memory such as ROM or EEPROM, allowing the circuit to be programmed to perform a specific function. Once programmed, the circuit can be used to implement a wide range of digital logic functions, including Boolean logic, arithmetic operations, and data manipulation.

PLAs are a type of field-programmable logic device (FPLD), which are used to implement custom digital logic circuits without the need for custom-designed integrated circuits. While PLAs are not as flexible as FPGAs, they offer a simpler and more cost-effective solution for many applications. They are commonly used in industrial control systems, automotive applications, and telecommunications equipment.

16.3 Hardware Description Languages

A hardware description language (HDL) is a specialized programming language used to design and describe digital circuits and systems. HDLs allow designers to create complex digital circuits by describing their behavior using a high-level language, rather than manually designing the circuit at the gate or transistor level.

VHDL (VHSIC Hardware Description Language) is one of the most commonly used HDLs. It was developed by the U.S. Department of Defense in the 1980s as part of the VHSIC (Very High-Speed Integrated Circuit) program. VHDL is a powerful language that can be used to describe the behavior of digital systems at multiple levels of abstraction, from the lowest level of gates and transistors to the highest level of system behavior.

In VHDL, digital circuits are described using entities, which represent the interface and behavior of a particular module, and architectures, which define the internal structure and behavior of the module. VHDL provides a wide range of constructs for describing the behavior of digital circuits, including conditional statements, loops, and procedures, as well as support for data types, signals, and ports.

VHDL is used extensively in the design and verification of digital systems, including ASICs, FPGAs, and other types of digital circuits. It is supported by a wide range of software tools, including simulators, synthesis tools, and verification tools, which allow designers to simulate, test, and verify their designs before they are implemented in hardware.

Chapter 17

Sequential Digital Systems

17.1 Digital Memory

All the previous circuits were "combinational" circuits. This means that identical inputs will always produce identical outputs. These circuits are useful for straightforward applications such as a door bell, turning on a light, an alarm, ...i.e. simple circuits that don't need a memory. Another application is the implementation of mathematical function as in a calculator. They can however not be used for more complicated functions like event counting, or for state machines like processors: devices that have a state, i.e. some internal circuitry determines in which state they are.
Sequential Logic adds a new dimension to the use of digital circuits: they contain a memory. You can now work with circuits whose response is "event dependent", like electronic locks, counters or state machines: the way they behave not only depends on the current input, but also on the past and the inputs they received previously.

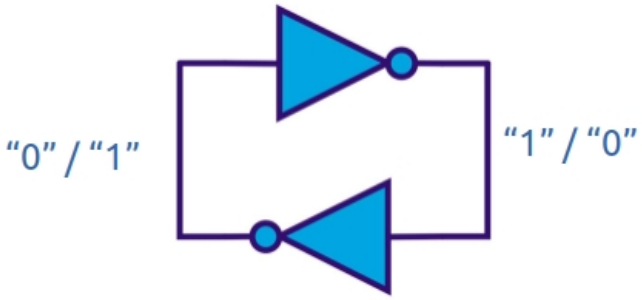


Figure 17.1

A simple memory element, capable of storing one bit of information, is shown in figure 17.1. It are essentially two NOT-gates back-to-back. If the input of one gate is "0" (LOW) the output will be "1" (HIGH). This signal is then propagated through the other NOT-gate to the input of the first one. In this way, the signal is sustained and the element remains in memory. This is also called a *flip-flop* or *latch*: a circuit that has two stable states that can store state information.
The problem however is that there is no mechanism to change the memory. This can be solved with an SR-flip-flop.

17.2 SR Flip-Flop

An SR flip-flop replaces the two NOT-gates from figure 17.1 with two NAND-gates as in figure 17.2. If input S ("Set") and R ("Reset") are both low, nodes X and Y are high. When one input of a NAND-gate is high, the gate acts as a NOT-gate on the other input: $\overline{1 \cdot A} = \overline{A}$. So to two NAND-gates with each an input at 1 acts as the connected NOT-gates in figure 17.1. This means that the bit present in memory Q is sustained. The other output is always not-Q: $\overline{Q}$. When S is high and R is low, node X is low and Q becomes high, irrespective of

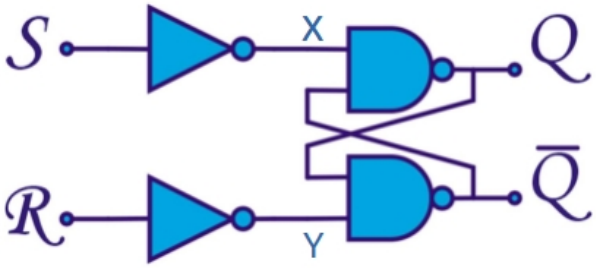


Figure 17.2

what the memory state was. Consequently, the other NAND-gate has two high inputs and produces $\overline{Q} = 0$. When R is high and S is low, $\overline{Q}$ becomes high and $Q = 0$.
It is not allowed to make both inputs high: in that case, both Q and $\overline{Q}$ become high. After S and R return to low, there is a conflict: only one of them can stay high in the end. This will be the one for which the signal propagates the fastest through the different gates. This is hard to predict, so we don't know what will be the memory state of the latch - an unacceptable situation.
The truth table is shown in figure 17.3. Note how the first line means that memory state Q is unchanged. "NA" stands for "Not Allowed".
The symbols for clocked and unclocked SR flip-flop are represented in figure 17.4. The difference is that a clocked flip-flop has an additional input for a clock signal: a block signal that alternates between low and high with a constant period, for instance generated by a relaxation oscillator from chapter 11.5. A transition of the memory state will only happen at a rising edge of the clock signal, i.e. only the values that S and R have at the time of the rising edge matter. In an unclocked flip-flop, the output changes whenever S or R are set to high. The use of a clock synchronizes the behavior of the flip-flop, and avoids problems when they are used in larger circuits. It is therefore recommended to always use the clocked version.

S	R	Q
0	0	Q
0	1	0
1	0	1
1	1	N/A

Figure 17.3

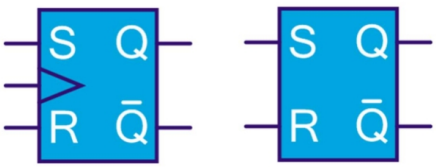


Figure 17.4

For the clocked version, we can also determine the *transition table*. It shows how the input

has to be set to obtain a specific output Q_{next} after the next transition, given the current (or present) output $Q_{present}$. For the SR flip-flop, the transition table is given in figure 17.5. An "X" entry means that the value can be either 0 or 1 - it's basically a "don't care". Convince yourself that this table gives the same results as the one in figure 17.3.

S	R	$Q_{Present}$	Q_{Next}
0	X	0	0
1	0	0	1
0	1	1	0
X	0	1	1

Figure 17.5

17.3 D-Latch

A D-latch is an adaptation of the SR flip-flop, as in figure 17.6. It works as follows:

- When input W is low, both nodes X and Y are high, so the NAND-gates act as interconnected NOT-gates and the existing memory bit Q is maintained.
- When W is high, node X is equal to $\overline{D}$ and $Y = \overline{X} = D$. So if $D = 1 \rightarrow X = 1 \rightarrow Q = 1$ and if $D = 0 \rightarrow Y = 0 \rightarrow \overline{Q} = 1$ and $Q = 0$, i.e. when $W = 1$, Q takes the value of the data bit D (so "D" stands for "Data").

Note how this topology avoids that $X = Y = 0$, the situation of the SR flip-flop that is not allowed. The symbol is shown in figure 17.7.

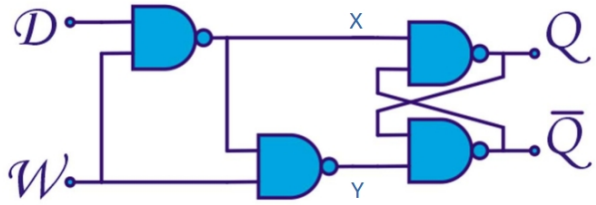


Figure 17.6

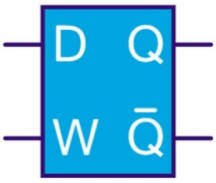


Figure 17.7

A very useful change to the D flip-flop is the clocked configuration, where input W is replaced by a clock entrance. Input D will only be transferred to the memory state Q on the rising edge of the clock signal - it is edge-triggered.

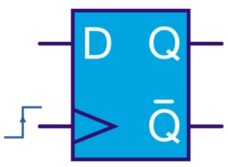


Figure 17.8

D	$Q_{Present}$	Q_{Next}
0	0	0
1	0	1
0	1	0
1	1	1

Figure 17.9

An edge-triggered D-flipflop is often used as a delay element in a digital filter, as in figure 17.10. The input signal propagates through the delay line and is moved one step to the right

at each rising edge of the clock. So the value $x(t)$ at time t at the input of the line is available after one period T at the output of the first D-flipflop, at the output of the second flipflop after 2 T, and so on.

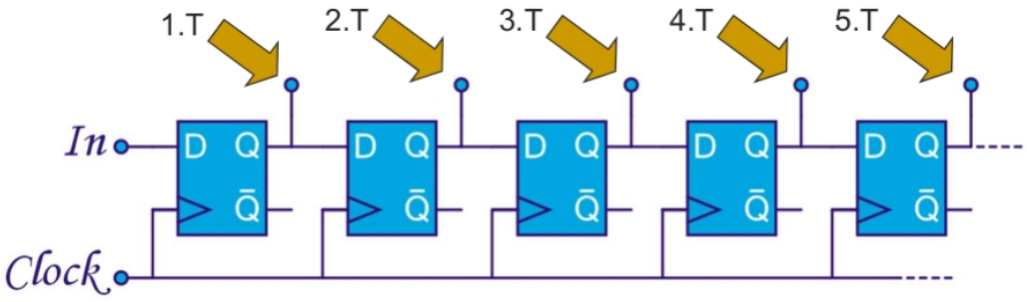


Figure 17.10

A finite impulse response (FIR) filter is a digital filter that computes an output signal $y[n]$ based on the current and previous values of input $x[n]$:

$$y[n] = a_0x[n] + a_1x[n-1] + a_2x[n-2] + \ldots + a_{N-1}x[n-N+1]$$

where N is the number of so-called tabs. For instance, if $N = 2$ and $a_0 = a_1 = 1/2$, we have an averaging filter of order 1, which removes the higher frequency components in the signal. All $x[n-k]$ tabs are supplied by a D-flipflop delay line. Digital filters will be studied in the course on Signals & Systems (ES311).

17.3.1 Implementation of the Edge-trigger

To implement the D-flipflop with an edge trigger, we use the circuit from 17.11. The part at the left is the *master*, the part connected to the output is the *slave*. Both can operate as a memory element, and each does this during part of the clock cycle.

Their operation is controlled by NOT-gates with inhibitor gates (14.3.1) so they behave

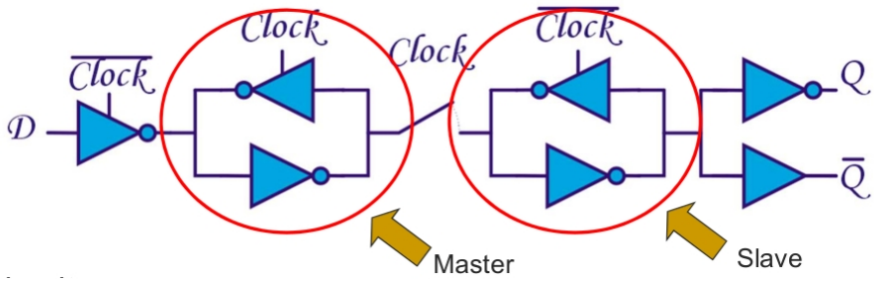


Figure 17.11

differently during different parts of the clock cycle:

1. When the clock is low:
   The circuit can be reduced to the one in figure 17.12. The master is connected to the input and follows the input through two NOT-gates in series. He is disconnected from

Figure 17.12

the slave, which operates as a memory element with two active NOT-gates back-to-back, and retains thus the value it has.

2. When the clock is high:
When the clock signal goes from low to high (i.e. the rising edge) the master is disconnected from the input and connected to the slave. The master is transformed in a memory element and retains the value it saw at the input just before the rising edge. This value is transmitted through the slave, which in turn is now two NOT-gates in series, to outputs Q and $\overline{Q}$, as in figure 17.13.

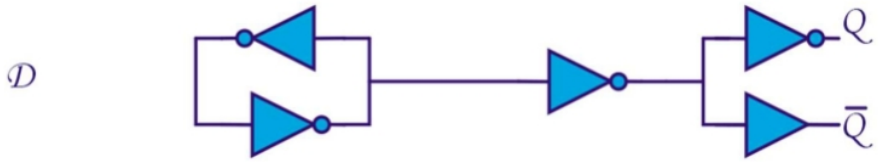


Figure 17.13

The consequence is that the data D that is present at the input at the rising clock edge, is transferred to the output and stays there for the entire clock cycle.
An SR flipflop can also be implemented with and edge trigger with the circuit in figure 17.14. When the clock is low, the initial part is an odinary (not edge-triggered) SR flipflop as we saw before. The slave is then just a memory element. When the clock goes high, S and R are disconnected from the circuit, and the inputs of the NAND-gates are connected to the supply (high) so the gates function as NOT-gates. The master becomes the memory element and its bit is transferred via the slave to the output.

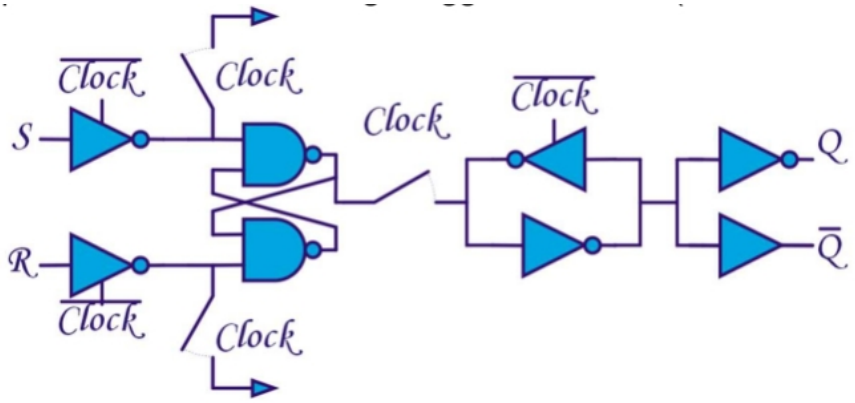


Figure 17.14

17.4 JK Flip Flop

The JK flipflop is an extension of the SR flipflop, where also the undefined input can be used. The consequence is that when the two input J and K are the low (either low or high), the memory state Q is retained, but when they are both equal to one, the memory state is flipped. The truth table is shown in figure 17.15. The symbol of this fliflop is exactly the same as for the SR flipflop, but with other names for the inputs - see figure 17.15. As before there is an clocked and unclocked version - obviously the clocked version is preferred.

J	K	Q
0	0	Q
0	1	0
1	0	1
1	1	$\overline{Q}$

Figure 17.15

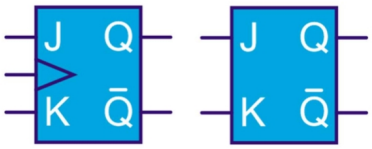


Figure 17.16

The transition table for the clocked version is shown in figure 17.17. For example, if we have a zero in memory and we want to keep it, as in the first line, we either keep the current memory with $J = 0, K = 0$, or we set the memory to zero: $J = 0, K = 1$. So to have this transition, J should be zero but hte value of K doesn't matter ("X" stands for "Don't care"). The reasoning for other lines is similar.
A possible implementation can be found in figure 17.18. For the first three lines of the truth table, there is no problem and we can just either generate the memory state with the master based on the input ("Set": $J = 1$","$K = 0$" or "Reset": $J = 0, K = 1$) or keep the value in memory $J = K = 0$. But if $J = K = 1$, we need to invert the value in memory, so we have to fetch it and invert it.

J	K	$Q_{Present}$	Q_{Next}
0	X	0	0
1	X	0	1
X	1	1	0
X	0	1	1

Figure 17.17

17.5 Applications

We present in this section two small applications where memory elements are used: a simple counter and a register. In the next section, we discuss the sequential circuit design method, which is a principled approach to implement (complex) sequential systems.

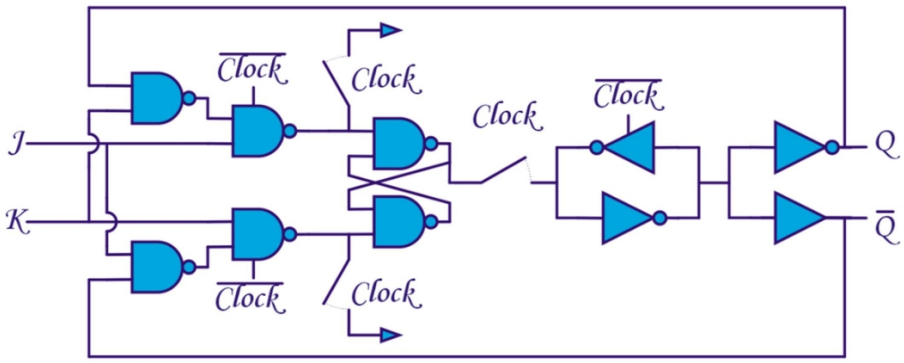


Figure 17.18

17.5.1 The Counter

A simple counter can be implemented with the circuit in figure 17.20. It is a counter for 4 bits, so it counts from 0 to $2^4 - 1 = 15$ in binary.
All flip-flops have initially a zero memory state. The J and K inputs for all flip-flops are both set to 1 which means that the memory value will be flipped at each rising edge at the clock port. The first flip-flop, which contains the least significant bit, flips every time the input signal goes from low to high. The output $\overline{Q}$ of this flip-flop is the clock input of the next one, so the second flip-flop flips when the first bit Q_0 goes from 1 to 0. This repeats for every two flip-flops and means that every flip-flop flips once for every two flips of the previous one.
The result is that the the output signal that is produced is an increasing sequence of binary numbers, as in figure 17.19, up until 15 and then the entire sequence starts again. This counter can easily be extended to higher numbers.

Q_3	Q_2	Q_1	Q_0
0	0	0	0
0	0	0	1
0	0	1	0
0	0	1	1
0	1	0	0
0	1	0	1
0	1	1	0
0	1	1	1
1	0	0	0
1	0	0	1
1	0	1	0
1	0	1	1
1	1	0	0
1	1	0	1
1	1	1	0
1	1	1	1

Figure 17.19

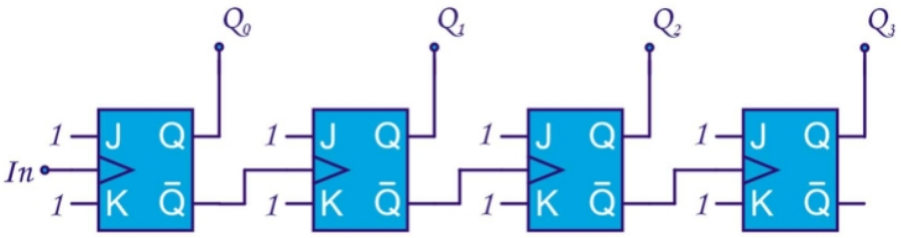


Figure 17.20

17.5.2 The Register

A register is similar to memory, but has typically more functions, i.e. it can work in several modes. The register we show here has 4 D-flipflops as memory cells and has 2 modes: parallel and series (or shift). The mode is set by a multiplexer such as we saw in section 14.2.1.
In parallel (P) mode the multiplexers put the inputs P_0 to P_3 directly on the D input pin of the latch. These values are thus read into memory whenever the W input is high.
In series (S) mode, the initial value S is read into the first flip-flop, and every other flip-flop changes its value to the one of the previous latch - when W is high, off course.
A register is not really used as a memory: it is lot smaller and can also perform small operations. The series (or shift) operation from the register on the right, for instance, corresponds to a (binary) multiplication by 2 when the initial value S is zero. A register is usually added to an ALU (Arithmetic-Logic Unit) to memorize small amounts of data that are needed for intermediary steps in calculations.

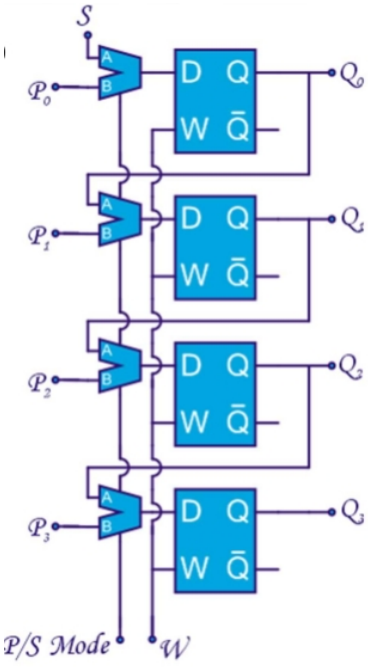


Figure 17.21

17.6 Sequential Circuit Design

A sequential circuit consists of three elements:

1. A flip-flop array, that contains a number of flip-flop memory cells and retains the state of the circuit.
2. I/O logic, a combinational circuit that generates the output based on the input and the memory.
3. the next-state logic, another combinational circuit that transforms the circuit input into the J/K, S/R or D inputs for the flip-flops used in the array such that the circuit will evolve to the next state.

The memory state makes the analysis of a sequential circuit more difficult than a combinational circuit. To perform this analysis, we rely on a *state diagram*.

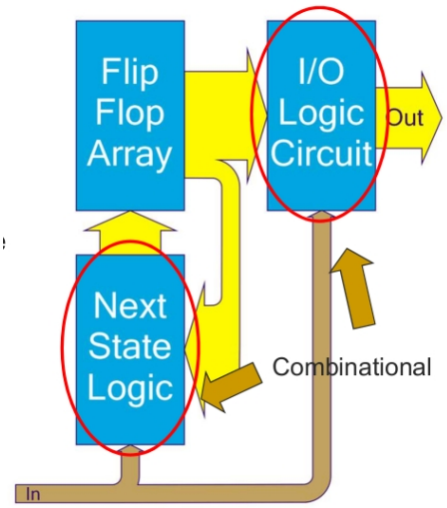


Figure 17.22

A state diagram represents the state a circuit is in, the possible inputs (if there are any) and where to system will evolve to, based on the present state and the input - i.e. what will be the next state. We represent the states of the system as circles containing the state name, and the transitions as arrows between states. An arrow is associated with the current input and the output of that transition. Figure 17.23 shows the state diagrams for SR, JK and D

flip-flops. Obviously, each flip-flop has 2 states ($Q = 0$ and $Q = 1$).

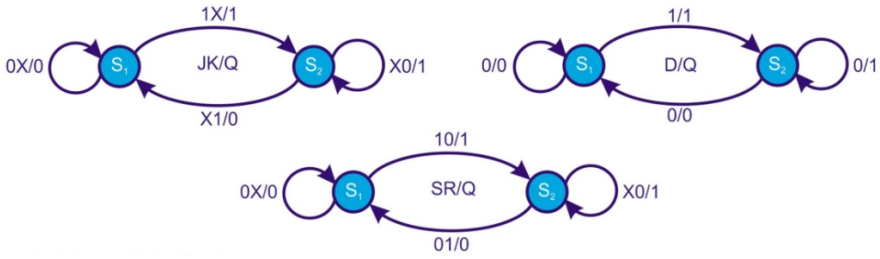


Figure 17.23

Some other examples are:

1. **Sequence detector** (figure 17.24)
   This circuit detects a specific sequence in an input bitstream. Specifically, the detected

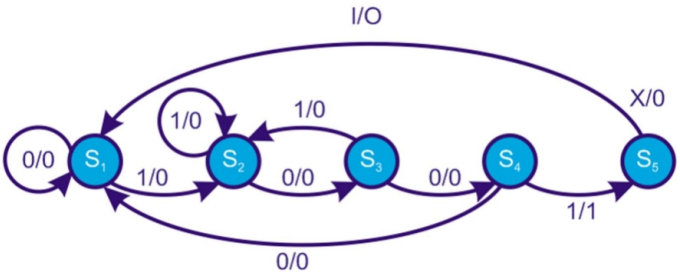


Figure 17.24

   sequence in figure 17.24 is "1001". State S_1 is the start state, S_2 is the state after observing a 1, S_3 the state after "10" and so on. If a bit that is not part of the sequence is observed, two things can happen: if this bit is a 0 the system returns to start state S_1, but if it is a 1, it can be the start bit of a new "1001" sequence and the system moves to state S_2. After a sequence "1001" is observed, the output goes high (see the transition from S_4 to S_5) and the system returns to S_1.
   Because there are 5 states, we need 3 flip-flops to represent them.

2. **Binary Coded Decimal (BCD) counter** (figure 17.25)
   This is a counter that goes from 0 to 9 and then resets to 0. It requires no input and the output is the value of the next state. This choice of output avoids additional logic to transform the state value to the output. An extension is an up-and-down counter, where a input 0 keeps counting up, and a input 1 counts down as in figure 17.26. Because there are 10 states, we need 4 flipflops.

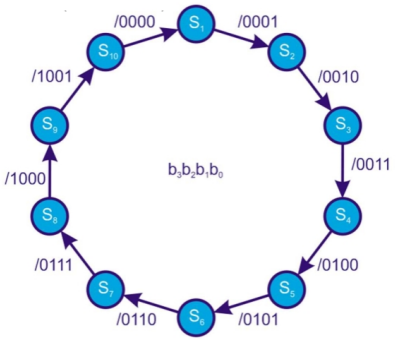


Figure 17.25

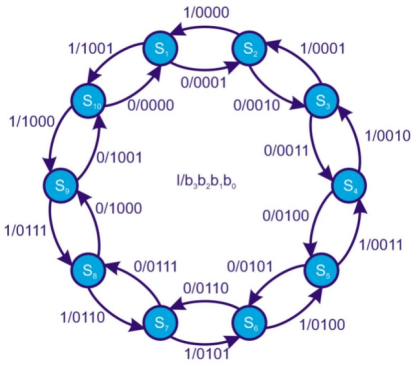


Figure 17.26

17.6.1 Example: An SR-FF using a D-FF

We will walk through an example to illustrate the sequential design procedure. The goal is to make an SR flip-flop based on a D flip-flop.
The design procedure goes as follows:

1. Draw the state diagram, as below.

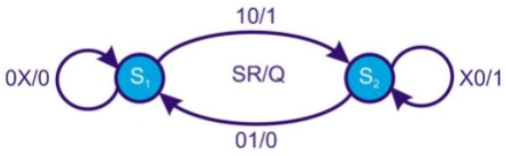


Figure 17.27

2. Derive the transition ("Present-Next state") table. The third column should be interpreted as: if I'm in state S_1, where do I move to with inputs S and R. Same goes for the last column. From this table, we determine the number of flip-flops we need. The

S	R	S_1	S_2
0	0	S_1	S_2
0	1	S_1	S_1
1	0	S_2	S_2
1	1	X	X

Figure 17.28

   answer is 1 because there are only 2 states.

3. Define for each state the state F_i of the flip-flop and the output Q generated from F_i. Whenever possible, we choose $Q = F_i$. The states of the D flipflop are either 0 or 1.

We match the outputs Q of the resulting SR FF directly to these states.

State	F_1
S_1	0
S_2	1

Figure 17.29

State	Q
S_1	0
S_2	1

Figure 17.30

4. The next step is a complete state table. This table contains all required information:
   (a) the possible inputs S and R,
   (b) combined with the different possible (present) states $F_{1,p}$,
   (c) the different 'next' states $F_{1,n}$ resulting from these states and inputs,
   (d) the settings of the input of the D-FF to go from the present to the next state.

S	R	F_1|p	F_1|n	D
0	0	0	0	0
0	0	1	1	1
0	1	0	0	0
0	1	1	0	0
1	0	0	1	1
1	0	1	1	1
1	1	0	X	X
1	1	1	X	X

Figure 17.31

5. We construct the Karnaugh table to generate the combinational circuit to produce the input D of the D-flipflop. The inputs are S and R and the present state $F_{1,p}$. The resulting equation is $D = S + \overline{R} \cdot F$.

D - Table

F_1 \ SR	00	01	11	10
0	0	0	X	1
1	1	0	X	1

$$D = S + \overline{R}F$$

Figure 17.32

6. Based on this expression, we can generate the circuit, with the required flipflops, as in figure 17.33.

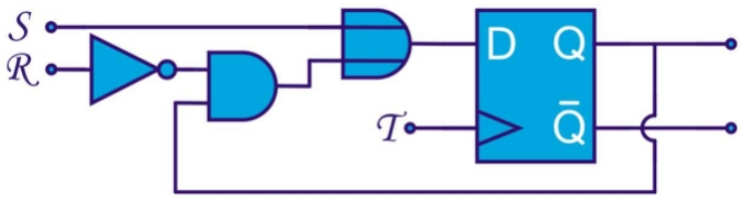


Figure 17.33

17.7 Memory Types

We give a brief overview of memory types found in modern hardware devices.

17.7.1 Static RAM

Static RAM (SRAM) is a type of computer memory that uses a flip-flop circuit to store each bit of data. SRAM is called "static" because it doesn't need to be periodically refreshed like dynamic RAM (DRAM). This means that SRAM is generally faster and more reliable than DRAM.
SRAM is commonly used as a cache memory in computer systems. It is also used in certain specialized applications where high speed and low power consumption are critical, such as in aerospace and military systems.
One of the key advantages of SRAM is its speed. SRAM can access data in just a few nanoseconds, which makes it ideal for applications that require high performance. However, SRAM is also more expensive than DRAM and requires more power, which limits its use in some applications.
Another advantage of SRAM is its durability. Since each bit of data is stored using a flip-flop circuit, it remains in place as long as power is applied to the memory chip. This makes SRAM ideal for applications that require data to be stored for long periods of time, such as in embedded systems or digital signal processing.
Overall, SRAM is a powerful and flexible type of computer memory that is used in a wide range of applications where high speed, low power consumption, and durability are important factors.

17.7.2 Dynamic RAM

Dynamic RAM (DRAM) is a type of computer memory that stores each bit of data as a charge on a capacitor within an integrated circuit. DRAM is called "dynamic" because the charge on each capacitor gradually leaks away over time, and needs to be periodically refreshed to prevent data loss. This means that DRAM is generally slower and less reliable than SRAM.
DRAM is the most common type of memory used in modern computers. It is used for main memory, which is the memory that the computer uses to store the programs and data that are currently being used.
One of the key advantages of DRAM is its cost. DRAM is cheaper than SRAM and requires less power, which makes it ideal for applications that require large amounts of memory, such as personal computers.
However, DRAM is also slower than SRAM, with access times typically in the range of tens

of nanoseconds. This means that it is less suitable for applications that require high performance, such as in embedded systems or digital signal processing.
Another disadvantage of DRAM is its volatility. Since each bit of data is stored as a charge on a capacitor, the data is lost when power is removed from the memory chip. This means that DRAM is not suitable for applications that require data to be stored for long periods of time without power, such as in embedded systems or digital signal processing.
Overall, DRAM is a cost-effective and widely used type of computer memory that is ideal for applications that require large amounts of memory at an affordable price.

17.7.3 Read-Only Memory

Read-only memory (ROM) is a type of computer memory that is pre-programmed with data that cannot be modified or changed. This type of memory is commonly used for storing firmware and other low-level system information that is critical for the proper functioning of a computer or other electronic device.
There are several different types of ROM, including PROM, EPROM, and EEPROM. Here are the key differences between these three types:

- PROM (Programmable Read-Only Memory): This type of ROM is programmed once by the manufacturer, using a special device called a PROM programmer. Once programmed, the data in a PROM chip cannot be changed or erased. PROMs are commonly used for storing firmware and other fixed data that is unlikely to change.
- EPROM (Erasable Programmable Read-Only Memory): This type of ROM can be erased and reprogrammed using ultraviolet light. EPROM chips have a small window on top that allows the ultraviolet light to reach the memory cells. To erase an EPROM chip, it must be exposed to UV light for several minutes. Once erased, new data can be programmed into the chip using a PROM programmer. EPROMs are commonly used for development purposes and in systems where the firmware may need to be updated.
- EEPROM (Electrically Erasable Programmable Read-Only Memory): This type of ROM can be erased and reprogrammed electronically, without the need for UV light. EEPROMs are commonly used for storing small amounts of data that may need to be updated periodically, such as configuration settings or user preferences. EEPROMs are slower than other types of ROM and have a limited number of write cycles, meaning that they can only be reprogrammed a certain number of times before they wear out.

Chapter 18

A/D and D/A Converters

18.1 Introduction

A *converter* is an electronic device that transforms a signal:

- From an analog to a digital signal (ADC)
- From a digital to an analog signal (DAC)

An *analog* signal is continuous over time, i.e. it has a value for every time t, and can take any value (typically between a minimum and maximum value).
A *digital* signal can only be equal to a certain number S_i, with usually $0 \leq S_i \leq 2^n - 1$. S_i can thus be stored as a binary number of n bits. The number S_i is related to the value V_i of the sample: $V_i = (V_{max} - V_{min})\frac{S_i}{2^n}$. The distance between two successive levels is one LSB or least significant bit.
Furthermore, a digital signal is not continuous over time: it represents a signal at specific moments in time. The period between these time instants is usually constant and equal to the sampling period T_s, such that sample k is taken at time $k\,T_s$. The sample rate ω_s is equal to $\frac{2\pi}{T_s}$. To reconstruct an analog signal perfectly from a sampled digital version, the sample rate must be at least twice the maximal frequency present in the signal. This is the theorem of Shannon: $\omega_s > 2\,\omega_{max}$.
Figure 18.1 represents the continuous analog signal V_{in} in red, the samples as red dots with sample period T_s and the allowed values of S_i between V_{min} and V_{max}.
An *Analog-to-Digital Converter* (ADC) will convert an analog signal to discrete voltage levels at specific, discrete moments in time. The latter is characterized by the sampling period T_s. The reduction to discrete values leads to a quantization error ϵ_s. Preferably, this error should be smaller than the resolution of the measurement. The sampling rate must be higher than the Shannon rate $2\,\omega_{max}$. If ω_{max} is unknown or too high (noise!), a low-pass (anti-aliasing) filter must be used.
The goal of a data converter is to transfer a signal from the real world to a digital storage device. The signal is picked up by a sensor (an antenna, a microphone, a camera, ...) and will first be sampled by a sample-and-hold circuit, where the present value of the signal is stored on an analog memory (like a capacitor) and then converted by the ADC to a number of bits (a digital *word*) that can be transferred to a digital device for processing and storage. This pipeline is shown in figure 18.2.
This process can also work in the other way: a digital system (like a PC or microcontroller)

Figure 18.1

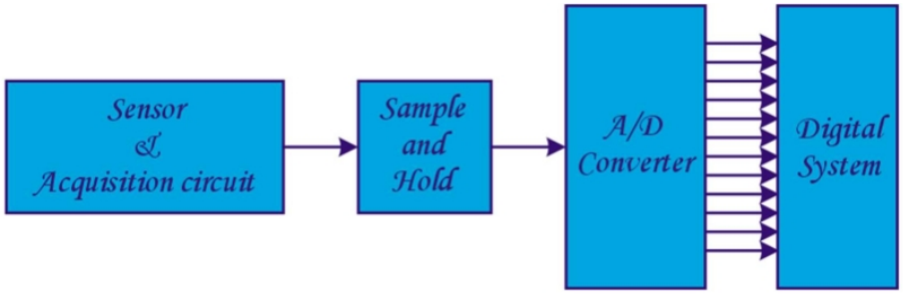


Figure 18.2

generates a digital actuator signal. This signal is then converted to an analog signal with a *Digital-to-Analog Converter* (and a filter to smooth out the transitions generated by the conversion process) and is then applied to an actuator.

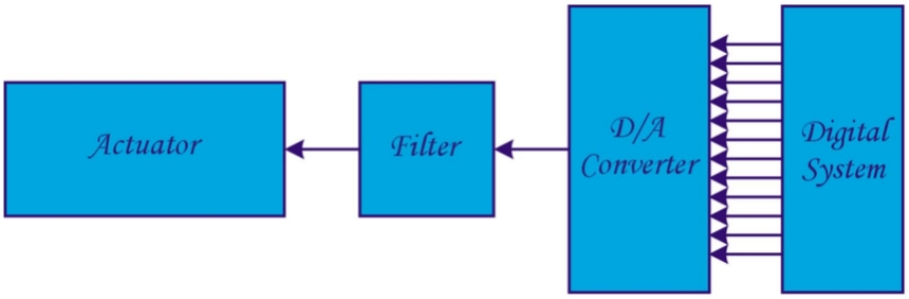


Figure 18.3

18.2 Characteristics of DAC & ADC

A DAC converts a binary word $b_0 b_1 \dots b_{n-1}$ of n bits to an analog output V_{out}:

$$V_{out} = \frac{V_{ref}}{2^n} \sum_{i=0}^{n-1} b_i 2^i$$

as in figure 18.4. The least significant bit (the resolution, or the spacing between two successive analog values) is equal to

$$LSB = \frac{V_{ref}}{2^n}$$

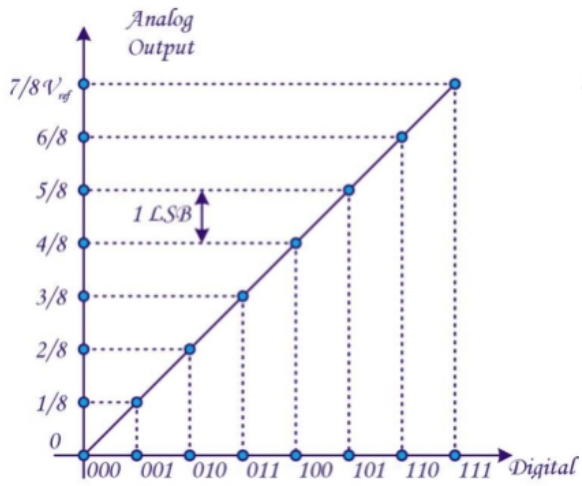


Figure 18.4

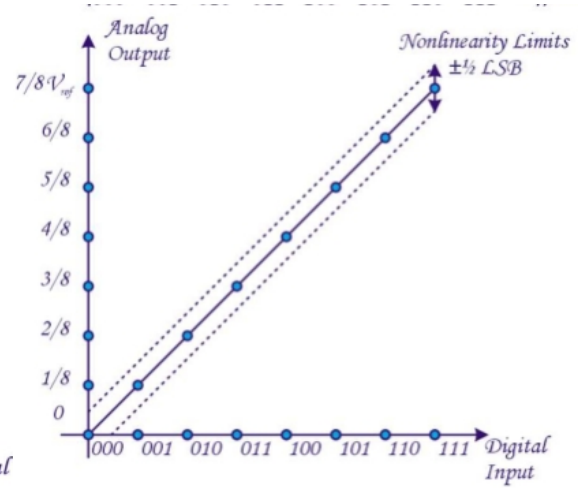


Figure 18.5

Linearity is important: we require that the error of the generated values is always limited to $\pm$ 1/2 LSB. If not, it is possible that a code word a that is lower than a code word b produces a higher analog value.

A real DAC contains many types or sources of errors beside non-linearity. A true output of a DAC is the red line in figure 18.6, which can be approximated by a best-fit line in blue. This line should keep the error with respect to the real line within 1/2 LSB. The offset error is the error due to the fact that the best-fit line doesn't go through the origin. Additionally, there is a gain error because the slope is not exactly 1.

More important are the deviations for each point between the measured line and the best-fit line. The *integral non-linearity* (INL) is the maximum deviation over the whole range, while the *differential non-linearity* (DNL) is the maximum deviation between two consecutive steps: we expect that for one step, the output moves by 1 LSB. The maximal deviation from this LSB is the DNL. The INL gives the linearity, while the DNL determines the resolution. Figures 18.7 and 18.8 show DACs with high INL and low DNL (left) and low INL and high DNL (right).

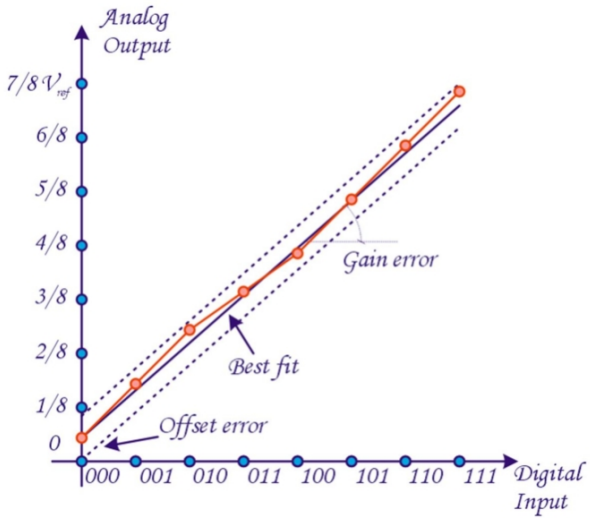


Figure 18.6

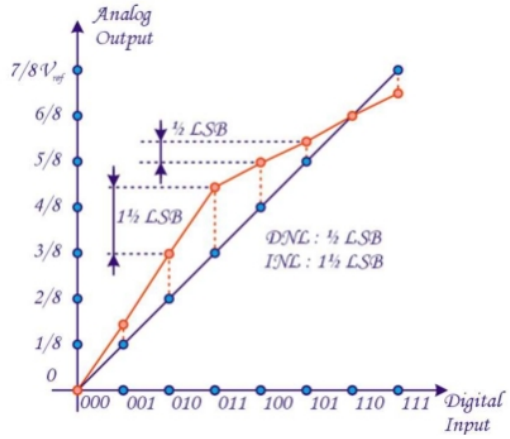


Figure 18.7

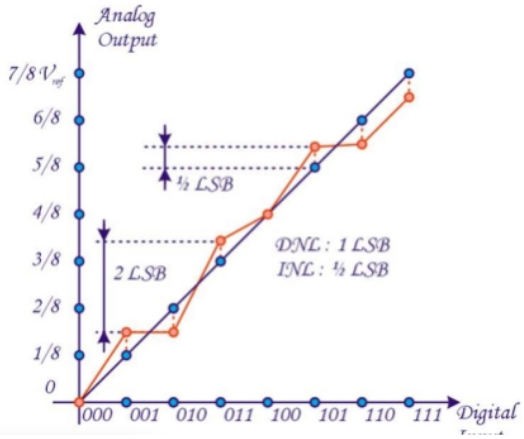


Figure 18.8

An ADC converts analog input value V_{out} into a binary word $b_0 b_1 \ldots b_{n-1}$ of n bits. An example of an ideal input-output relation with $n = 3$ is given in figure 18.9. It follows a staircase pattern where each codeword corresponds to a value of V_{in} between $\frac{k}{2^n} V_{ref} - \frac{1}{2}\frac{1}{2^n} V_{ref}$ and $\frac{k}{2^n} V_{ref} + \frac{1}{2}\frac{1}{2n} V_{ref}$ (except for 000). The spacing between consecutive codewords is 1 LSB, as required. The maximum error is $\pm$ 1/2 LSB.

As for the DAC, there is also the notion of INL (global error) and DNL (1-step error). For example, in the left part of figure 18.10, with blue the ideal curve and in red the real one, there is no input value that corresponds with the code "011", so the DNL is high. In the right figure, the relative error of each step is quite small, but there is a significant non-linearity, characterized by a large INL.

18.3 Sample-and-Hold Circuit

A sample-and-hold circuit temporarily stores the value of an input signal such that it can be processed and converted by an ADC. It consists of a switch, a capacitor and an OPAMP as

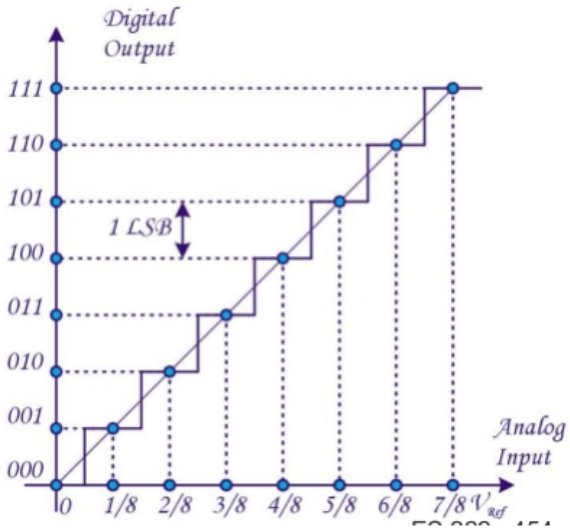


Figure 18.9

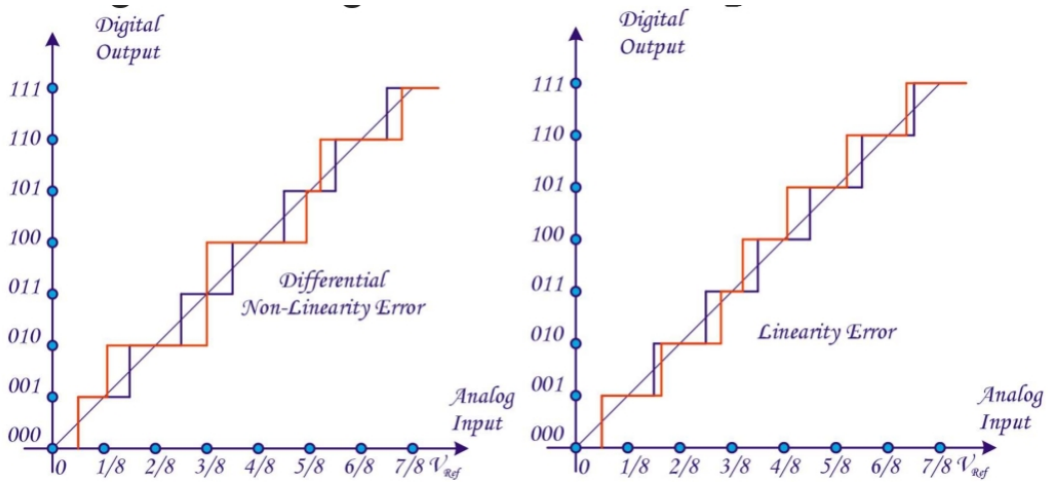


Figure 18.10

a unity gain buffer (see chapter 7.4.4), as in figure 18.11.

During the "Sample"-phase of the cycle (see figure 18.12, with in red the input signal and in blue the output of the sample-and-hold), the switch is closed. After a transition period t_A[1] the input of the OPAMP will follow the signal. The time t_A is the *acquisition time*. During the "Hold" part of the cycle, the switch opens and the capacitor will retain the value of the signal at the end of the sample period. During this time, the output $V_{S/H}$ is constant and the ADC can do the conversion. When the input impedance of the OPAMP is not infinite, a small current will be drawn and the voltage on the capacitor will slowly decrease: $\Delta V = \frac{I_{leak}}{C}\Delta t$; this is the *droop rate*.

[1]Because the OPAMP and the RC circuit formed by switch and capacitor are a higher order system, there is a settling time and a ripple.

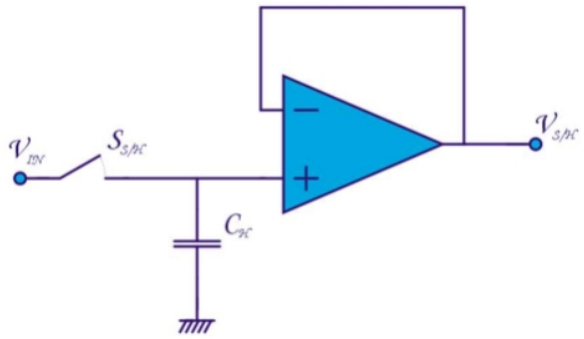


Figure 18.11

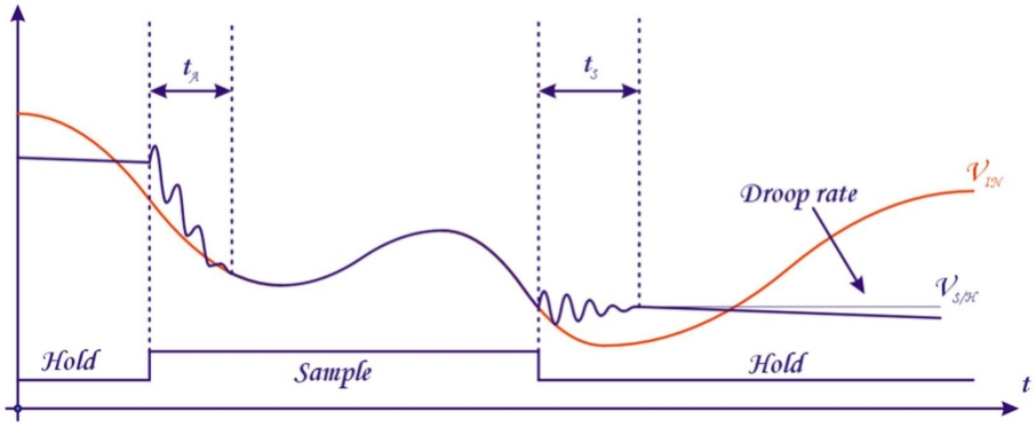


Figure 18.12

18.4 Digital-to-Analog Converters

In this section, we discuss 2 families of DAC's, the first one based on voltage distribution, the other one based on charge distribution. There is a trade-off between precision and speed of conversion.

18.4.1 Voltage Distribution

This converter uses a resistor network to generate the different possible output voltages, as in figure 18.13. The lowest voltage is $\frac{1}{2}\frac{V_{ref}}{2^n}$, the second level is $\frac{3}{2}\frac{V_{ref}}{2^n}$, and so on. This corresponds to the staircase pattern from figure 18.9. The voltage that is set at the the output is determined by a single switch that is closed while all the other are open. Since each voltage level requires its own switch, the number of switches is 2^n. To control the switches, the input word $b = b_1 b_2 \dots b_n$ is converted in a one-hot code as in figure 18.14.
This is a simple solution that works fine and the conversion can happen very fast (a single clock cycle is enough) for a small number of bits. But if we want a 12-bit decoder, we would need $2^{12} = 4096$ switches and a decoder network that transform a word of 12 bits into a switch setting containing 4096 bits, which is far from easy.

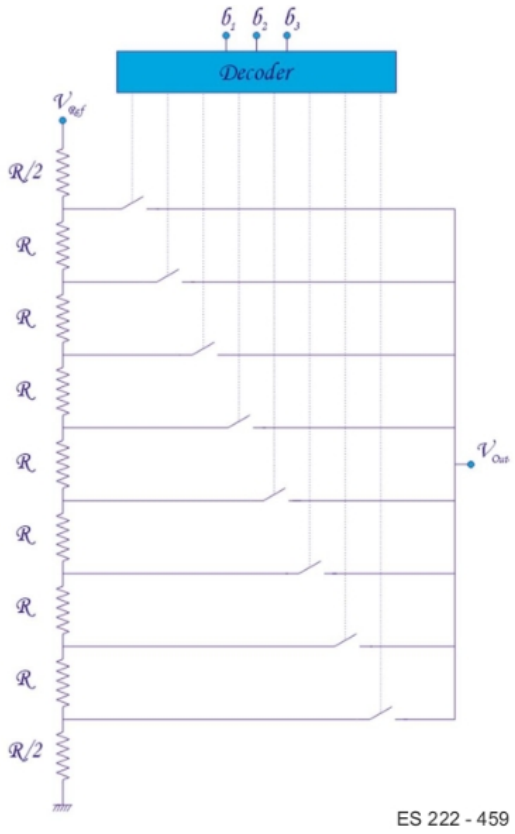


Figure 18.13

Binary Word $b_3b_2b_1$	Switch setting
000	00000001
001	00000010
010	00000100
011	00001000
100	00010000
101	00100000
110	01000000
111	10000000

Figure 18.14

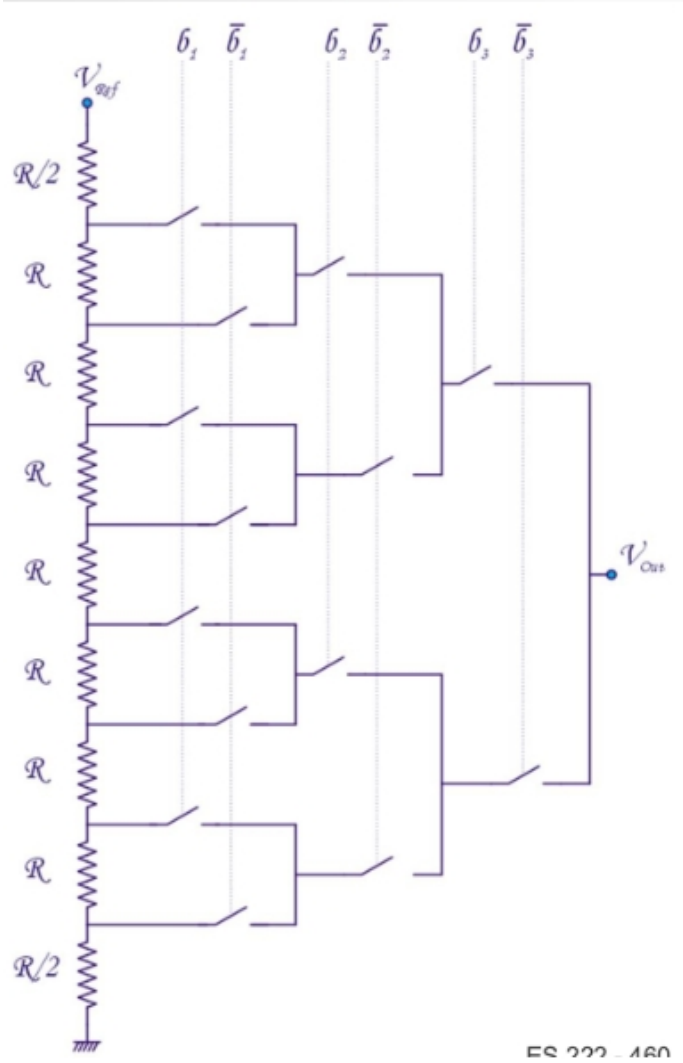


Figure 18.15

The complex decoder network for the switch settings can be avoided by implementing a switching network as in figure 18.15. Note how each bit either opens a one switch or the one below it, because these switches are controlled by b_i or $\overline{b_i}$, respectively.

For every codeword b, there will be only one conducting path of closed switches from the resistor network to the output. For instance, if $b = 000$, only in the bottom path are all the switches closed, so $V_{out} = \frac{1}{16}V_{ref}$. If the least significant bit b_1 switches from 0 to 1, the lowest switch opens but the one just above closes, so now there is a path (and only one path) between $\frac{3}{16}V_{ref}$ and V_{out}. If $b_1 = b_2 = b_3 = 1$ there is a path from voltage $\frac{15}{16}V_{ref}$ to V_{out}.

This configuration removes the complexity of the decoder network but requires $\sim 2^{n+1}$ switches.

An elegant solution when a high number of bits is required, is an R-2R-ladder, as in figure 18.16. Every switch S_i is associated with a bit b_i from the codeword. If the bit is zero, the switch is to ground, and if $b_i = 1$, S_i is at V_{ref}.
Assume we want to convert "0001". This means that the three top switches are connected to ground, and the bottom one is at V_{ref}. This situation is depicted in figure 18.17 (left). We take the part in the blue square and compute the Thévenin equivalent. Two resistances $2R$ in parallel give $R_{th} = R$, and $E_{th} = \frac{2R}{2R+2R}V_{ref} = \frac{V_{ref}}{2}$, as in the middle part of the figure. With R and R in series, these resistances form one resistance of value $2R$. If we want to transform the blue square in the middle figure, we perform the same calculations and obtain $R_{th} = R$ and $E_{th} = \frac{V_{ref}}{4}$. This process continues, and in the end we find $V_{out} = \frac{V_{ref}}{16}$.
If the codeword is "1000", only the top switch is connected to V_{ref} and all the other ones are at ground. All the resistors at the bottom form parallel combinations of $2R$ with $2R$, which equals R. If we keep doing this, we finally find that V_{out} is the output of a voltage divider with two resistances of value $2R$ between V_{ref} and ground, so this means $V_{out} = \frac{2R}{2R+2R}V_{ref} = \frac{V_{ref}}{2}$.
Try this analysis yourself with the codeword equal to "1010" and see how different voltage levels are added at V_{out}.
This implementation requires only n switches. The difficulty is creating the resistances with the right values throughout the whole ladder.

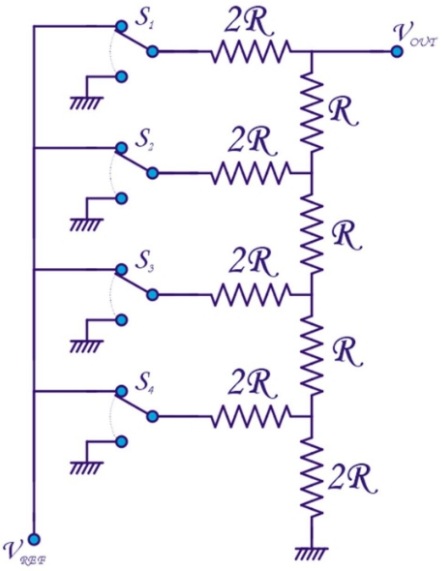


Figure 18.16

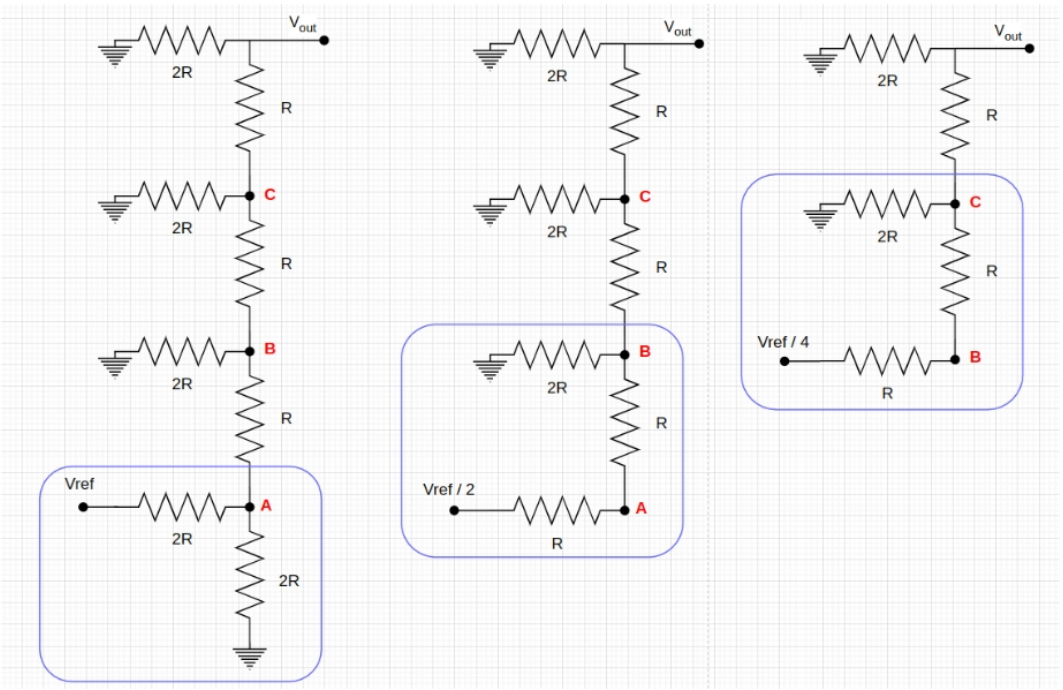


Figure 18.17

18.4.2 Charge Distribution

The DAC's from the previous section all worked by creating the different voltage levels are transforming them via a network of switches ot the output. The DAC's we'll study now accumulate charges over a network of capacitors to create the required voltage at the output. To implement this, we use the circuit from figure 18.18, with 5 capacitors in parallel. Because there are 4 switches S_1 to S_4, this is a 4-bit DAC.

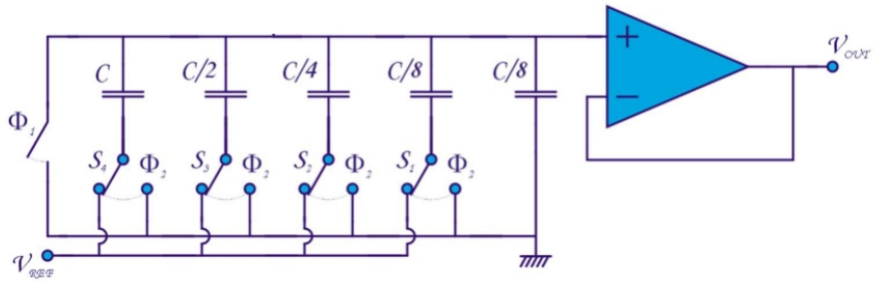


Figure 18.18

During the first phase of the cycle, all switches are closed i.e. connected to ground, so the different capacitors are all discharged. During the second phase of the cycle, switch S_i is connected to V_{ref} if $b_i = 1$ and to ground if $b_i = 0$. The equivalent capacitance C_{eq} from V_{ref} to the positive input of the OPAMP is:

$$C_{eq} = \sum_{i=1}^{n} b_i \frac{C}{2^{n-i}} = \frac{C}{2^n} \sum_{i=1}^{n} b_i 2^i$$

Figure 18.19 on the right shows the simplified circuit, with V_{out} as the result of voltage divider with C_{eq} and $2C - C_{eq}$ (because $C_{tot} = 2C$). So:

$$I = j\omega C_{eq}(V_{ref} - V_{out}) = j\omega(C_{tot} - C_{eq})V_{out}$$

$$\Rightarrow V_{out} = \frac{C_{eq}}{C_{tot}} = \frac{V_{ref}}{2^{n+1}} \sum_{i=1}^{n} b_i 2^i$$

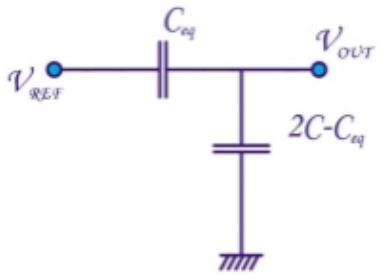


Figure 18.19

This means that every bit equal to 1 contributes a fraction of V_{ref} to the output voltage V_{out} in proportion to its value 2^i, just as required.

In practice, it can be difficult to make such small capacitances with high enough precision - knowing that the maximum error we may make is one LSB. A potential solution to this problem is the use of a shunt capacitance C_{shunt} as in the 6-bit converter of figure 18.20. At the right of C_{shunt}, the total capacitance we see is $C+C/2+C/4+C/4 = 2C$, but we actually

want to see $C/4$ from the left of C_{shunt}, looking to the right. With C_{shunt} in series with the total right capacitance $2C$ and the expression for two capacitances in series, we can compute C_{shunt}:

$$\frac{C}{4} = \frac{C_{shunt}\ 2C}{C_{shunt} + 2C}$$

with result $C_{shunt} = \frac{8}{7}\frac{C}{4} = \frac{2}{7}C$ as indicated in the figure. This solution, maybe repeated several times, allows to make precise converters with a resolution of several bits. Note however that each conversion is a two-step process: first the capacitors are discharges, and after that only the capacitors with bits equal to 1 are connected to V_{ref}.

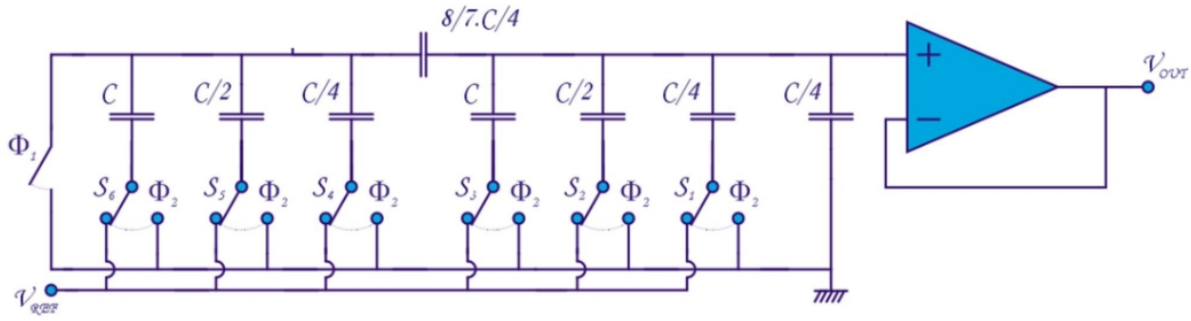


Figure 18.20

18.4.3 Serial Charge Distribution

Instead of having a capacitor for each bit, we can also charge a single capacitor every time a bit is high. The circuit used for this digital-to-analog conversion is shown in figure 18.21.

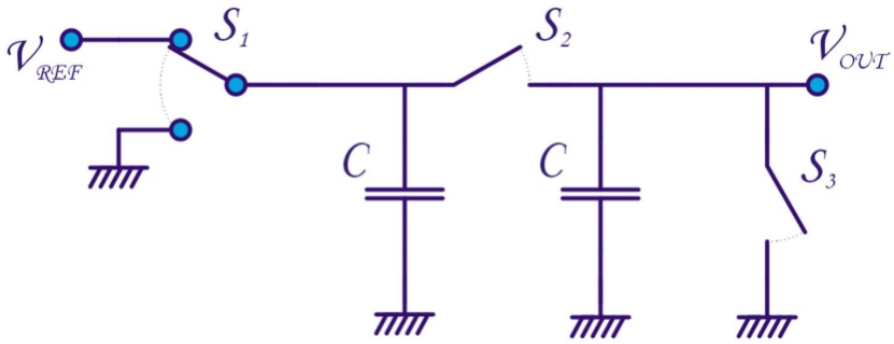


Figure 18.21

Initially, all switches are closed (with S_1 connected to ground) so that all capacitors can discharge. Then the conversion happens in n cycles for each of the b_i bits:

1. With switch S_2 open, switch S_1 is set to ground if $b_i = 0$ or to V_{ref} if $b_i = 1$. The charge on the first capacitor C is then $b_i\ C\ V_{ref}$, irrespective of the charge that was on it before.

2. Switch S_1 is reopened - this means it connects to neither ground or V_{ref}.

3. Switch S_2 is closed and reopened. This means that the charge that was present of the first and second capacitor is now equally distributed over both. So the charge on the second capacitor is now:
$$Q_{new} = \frac{Q_{old} + b_i C V_{ref}}{2}$$

4. This procedure continues until the last bit is evaluated.

So after n cycles, the total charge on the second capacitor is $Q = \sum_{i=1}^{n} \frac{b_i C V_{ref}}{2^i}$ and so

$$V_{out} = Q/C = \frac{V_{ref}}{2^{n+1}} \sum_{i=1}^{n} b_i 2^i$$

which is exactly what we need. This DAC can be easily constructed and can be very precise, but it will be slow because it requires $n+1$ cycles (including the discharge) to convert a word of length n.

18.5 Analog-to-Digital Converters

18.5.1 Serial Converter

The circuit of a serial converter is shown in figure 18.22. The first OPAMP, with a constant input voltage V_{ref}, is an integrator: $i = -V_{REF}/R = (-V_o)j\omega C$ so

$$V_O(j\omega) = \frac{1}{j\omega RC} V_{REF} \Leftrightarrow v_o(t) = \int_0^{\Delta t} \frac{1}{RC} V_{REF}\, dt = \frac{1}{RC} V_{REF}\, \Delta t$$

This linearly rising signal is compared by the second OPAMP with the input signal (from the sample-and-hold) V_{IN}. As long as V_{IN} is higher than the ramp, the output is high, and once the ramp passes V_{IN}, the output goes low. The result is a block signal whose length is proportional to V_{IN}.
This signal is applied at an AND-gate with a clock signal of frequency $f = \frac{1}{T}$ at the other input. The output of the AND-gate will be the clock as long as the signal is high, and low when the signal is low. The result is a sequence of pulses whose number N is proportional to V_{IN}. These pulses are counted by counter to produce the binary out signal.
The length of the signal at the output of the comparator is thus NT, reached when V_{IN} is equal to the ramp signal, and subsequently

$$V_{IN} = NT \frac{V_{REF}}{RC}$$

with the LSB equal to $T\frac{V_{REF}}{RC}$. This is a very precise conversion; the most likely sources of error are a change in R and C because the circuit heats up, or drift of the clock. The disadvantage is that conversion takes 2^n clock cycles, which can be prohibitively large.

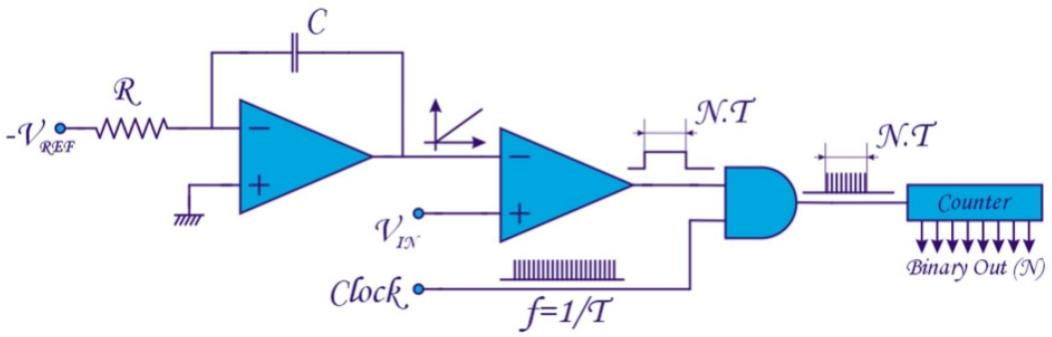


Figure 18.22

18.5.2 Double Integration

A solution to work even more precisely, is the double integration scheme from figure 18.23. At the input, there is switch that either places V_{IN} or $-V_{REF}$ at the input of the integrator. Initially, it will be V_{IN} and at the output of the integrator, we have a linearly decreasing signal whose slope is determined by the value of V_{IN}: $-\frac{V_{IN}}{RC}$. After N_{REF} clock cycles, the digital controller will put the switch at $-V_{REF}$.The signal at the output of the integrator is then $-\frac{V_{IN}}{RC}N_{REF}T + V_{th}$.
From that point on, the integration will be in the other direction, with a fixed sloped determined by V_{REF}, as in figure 18.24. After a time NT, the threshold voltage V_{th} is reached and the output of the comparator goes low. At that time, we have:

$$\frac{V_{IN}}{RC}N_{REF}T = \frac{V_{REF}}{RC}NT \Leftrightarrow N = \frac{1}{N_{REF}}\frac{V_{IN}}{V_{REF}}$$

In this way, the dependence on R, C and the clock period is removed. The LSB is $\frac{V_{REF}}{N_{REF}}$. The disadvantage is the same as before: conversion takes 2^n clock cycles and for some applications this can be too long.

18.5.3 Successive Approximation

A SAR or *Successive Approximation Register* performs a binary search over the space of codewords such that the corresponding voltage approximates the input voltage V_{IN}. The circuit is given in figure 18.25.

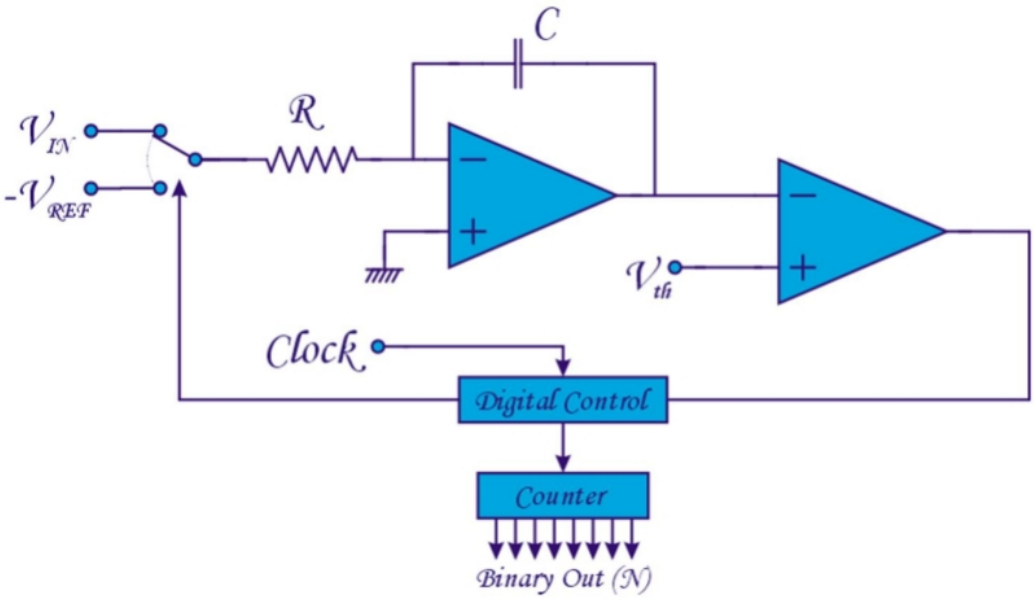


Figure 18.23

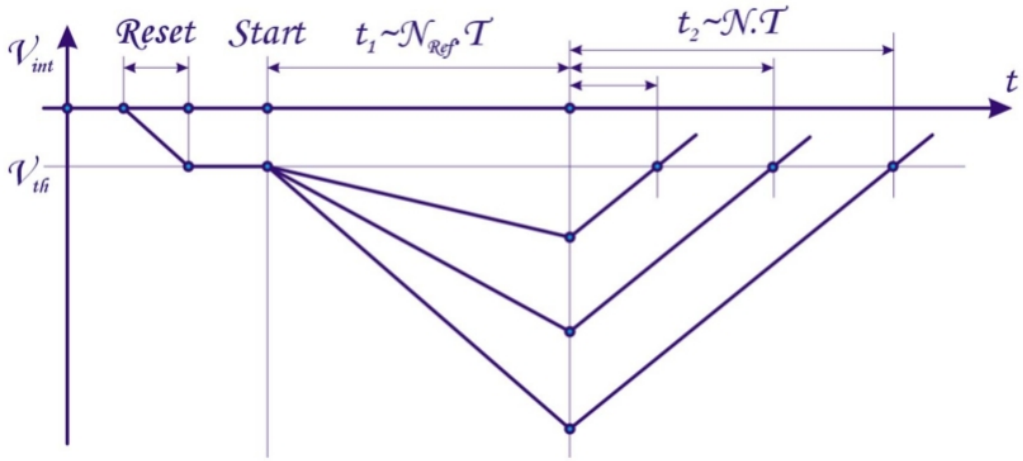


Figure 18.24

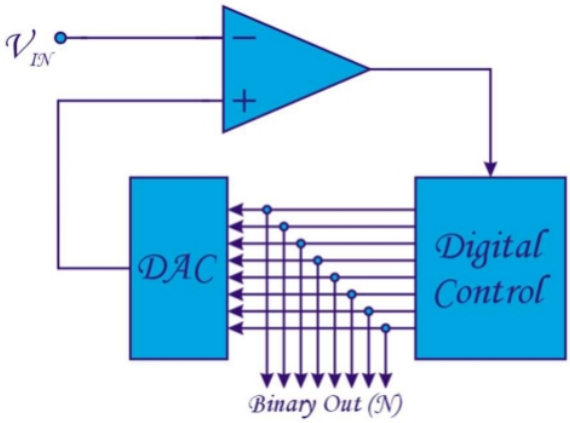


Figure 18.25

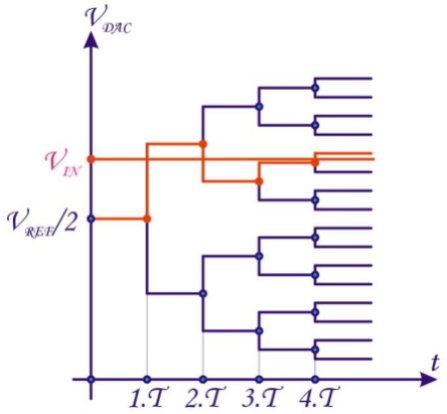


Figure 18.26

The digital controller generates a code that is transformed by the DAC into a voltage that is compared with the input V_{IN} by the comparator. Suppose the initial code was "1000". This generates a voltage in the middle of the domain, $\frac{1}{2}V_{REF}$. If the input signal is higher - as in figure 18.26, the output of the comparator is low and the controller knows the first "1" is correct. If the input were lower, the first bit (the most significant bit) must be a "0". During the next clock cycle, the digital controller will generate a code word "1100", the DAC will generate the corresponding voltage. This voltage is too high, as the figure shows, so the second bit will be a "0". This continues for the remaining bits. The correct codeword for this example is "10110".
This is a simple and straightforward converter that only needs n clock cycles to convert a signal to an n-bit word, but it requires a digital-to-analog converter and is (because of this DAC) less accurate.

18.5.4 Flash Converter

Figure 18.27 on the right shows the *flash converter*, so-called because it's a very fast analog-to-digital converter.
It works by generating all 2^n voltage levels between 0 and V_{REF} with a voltage ladder, just as the voltage distribution DAC from section 18.4.1. With 2^n comparators, the input voltage V_{IN} is compared with each voltage level. The output of all comparators at a voltage lower than V_{IN} will be one and the output of all those who lie higher will be zero. This is the so-called *thermometer code* that is transformed by the digital decoder network in a codeword of n bits.
This converter is very fast: it does the conversion in a single clock cycle, but it is not very accurate and requires a lots of components: 2^n resistors, 2^n comparators, each with an offset voltage that must be kept lower than 1 LSB, and a decoder network with 2^n inputs and n outputs. Construction of this converter is only feasible for $n \leq 10$.

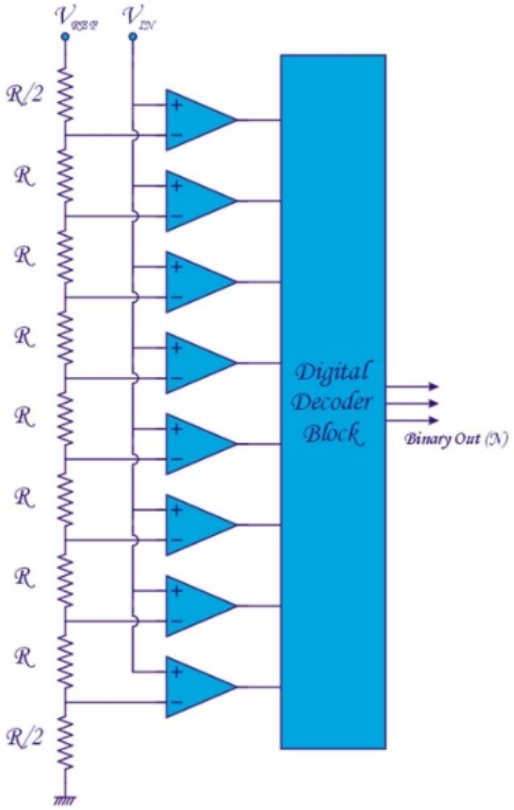


Figure 18.27

18.5.5 2-Step Flash

In stead of making a single N-bit flash converter, we can also make two consecutive $N/2$-bit converters, as in figure 18.28. The first ADC of size $N/2$ creates the first $N/2$ most significant bits of the output code word. These bits are then converted by a $N/2$ DAC to a voltage that is subtracted from the input signal. The range of this residual voltage is a single LSB of the first converter. A second ADC of size $N/2$, tuned to this range of 1 LSB, converts the residual

signal to an $N/2$ codeword. This word constitutes the $N/2$ least significant bits of the N-bit codeword. In this way, the complexity is reduced: from 2^N to $2 \times 2^{N/2}$.

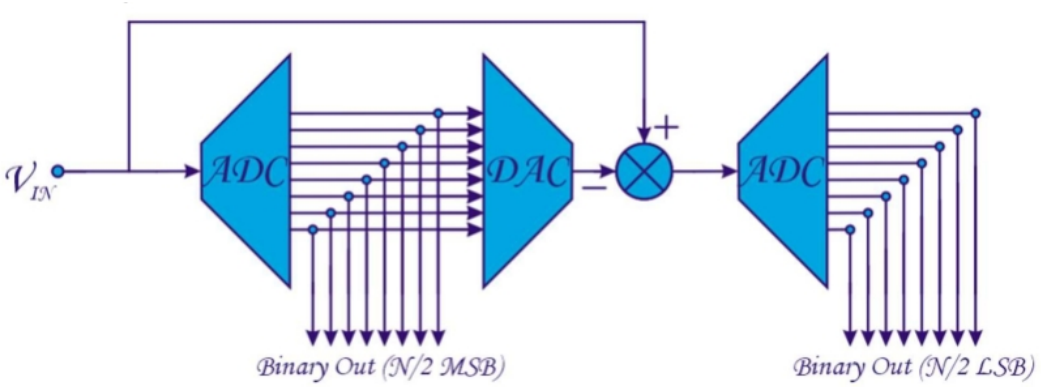


Figure 18.28

18.5.6 Pipeline Converter

A *pipeline converter* takes the idea from the 2-step flash converter even further: we convert a single bit at a time. Figure 18.29 shows a 3-bit pipeline. In the first comparator, we compare V_{IN} with V_{REF} - note that in this case, the range is 2 V_{REF}, so V_{REF} is exactly in the middle. MSB b_1 will be 1 if V_{IN} is higher and zero otherwise. To compute the residual - as in figure 18.28 - we don't use a DAC because with a single bit, we subtract either V_{REF} or ground from V_{IN} - hence the switch after the comparator, which is controlled by b_1. The next step would be to compare the residual with $\frac{1}{2}V_{REF}$. But this would require the generation of $\frac{V_{REF}}{2}$ and later $\frac{V_{REF}}{4}$, $\frac{V_{REF}}{8}$, etc .... To avoid this, we multiply the residual by two and compare with V_{REF} - an operation that gives the same result and avoids all these intermediary reference levels. This comparison gives b_2 and the same process continues.
This is a pipeline because the different stages can work independently from each other: after the first bit of sample 1 is converted, the conversion goes on but the first stage is now ready to convert the first bit of sample 2. So after n clock cycles, the first sample is completely converted, but after that there will be a new codeword after each clock period, because the samples move through the pipeline.

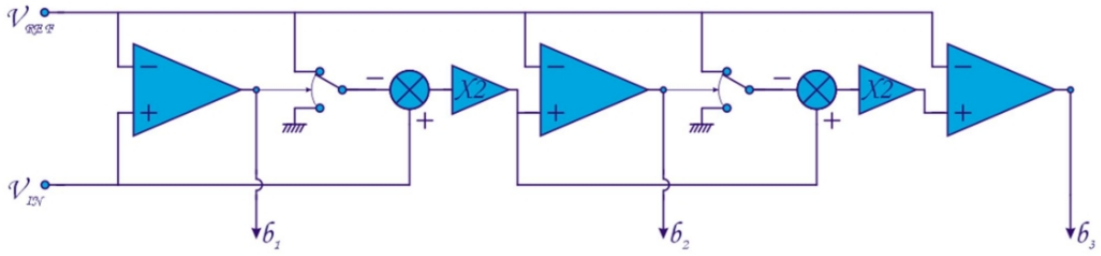


Figure 18.29

18.6 Additional Notes

In general, for both DACs and ADCs, we can state that the accuracy is inversely proportional to the speed: the more accuracy you require, the slower the conversion will be. For ADCs, we find:

- Serial converter: low speed, precision of $12 - 14$ bits.
- SAR converter: medium speed, $10 - 12$ bits.
- Flash converter: highest speed, $8 - 10$ bits

Anything over 14 bits bits will require special techniques, outside the scope of this course. A figure-of-merit for ADC is defined to compare converters with a single number:

$$FOM = \frac{P}{2^n f_c}$$

with P the power dissipation, f_c the conversion speed and n the number of bits. At this time, the best FOM $\sim 50nJ$.

Bibliography

[1] Alexander, Charles K., and Matthew Sadiku. **Fundamentals of Electric Circuits**. Boston: McGraw-Hill Higher Education, 2007.
Good reference for analysis of electric circuits.

[2] Kittel, Charles, and Paul McEuen. **Introduction to Solid State Physics**. John Wiley & Sons, 2018.
Classic reference book on solid-state physics.

[3] Sze, Simon M., Yiming Li, and Kwok K. Ng. **Physics of Semiconductor Devices**. John wiley & sons, 2021.
Classic reference book on electronic devices.

[4] Razavi, Behzad. **Fundamentals of Microelectronics**. John Wiley & Sons, 2021.
Modern reference book on microelectronics.

[5] Schilling, Donald L., Charles Belove. **Electronic Circuits: Discrete and Integrated.** New York: McGraw-Hill, 1989.
Another, older reference which overlaps with this course.

[6] Horowitz, Paul, and Winfield Hill. **The Art of Electronics.** Cambridge: Cambridge university press, 2002.
Focused on applications, with lots of tips & tricks for engineers.

Appendix A

Exam Questions

1. Discuss the PN-junction, including the small-signal model (4 and 6.2)
2. Discuss the BJT, including the small-signal model (5.1 and 6.5.1)
3. Discuss the MOSFET, including the small-signal model (5.2 and 6.5.2)
4. Discuss the biasing of the BJT (6.4.1)
5. Discuss the biasing of the MOSFET (6.4.2)
6. Discuss the four-resistor amplifier (7.1.2)
7. Discuss the common emitter/source amplifier (7.2.1)
8. Discuss the common basis/gate amplifier (7.2.2)
9. Discuss the common collector/drain amplifier (7.2.3)
10. Discuss the differential amplifier (7.3)
11. Discuss the operational amplifier, including sources of error (7.4)
12. What is the Miller capacitor? (8.2)
13. Discuss the class A amplifier (9.2)
14. Discuss the class B amplifier (9.3)
15. Discuss the Push-Pull amplifier (9.4)
16. Discuss the class C amplifier (9.5)
17. Discuss the class S amplifier (9.7)
18. Discuss the selective amplifier (9.8)
19. Discuss the impact of feedback (10)
20. Discuss the Wien Bridge oscillator (11.2)
21. Discuss the Colpitts oscillator (11.3)

22. Discuss the relaxation oscillator (11.5)
23. Discuss the Fantastron (11.6)
24. Discuss the diode rectifier (12.2)
25. Discuss the single voltage stabilizer (12.3)
26. Discuss the voltage stabilizer with transistor (12.3.2)
27. Discuss transistor-transistor logic (15.2)
28. Discuss emitter-coupled logic (15.3)
29. Discuss the Schmidt trigger (15.4)
30. Discuss the edge-triggered D-latch (17.3)
31. Discuss the sequential design procedure (17.6)
32. Discuss the R-2R DAC (18.4.1)
33. Discuss the DAC with charge distribution (18.4.2)
34. Discuss the serial ADC with single and double integration (18.5.1)
35. Discuss the SAR and Flash converter (18.5.3 and 18.5.4)