



Fundamentos de electrónica física y microelectrónica

J. M. ALBELLA

Profesor de Investigación del Instituto de Ciencia de Materiales.
Consejo Superior de Investigaciones Científicas
Madrid, España

J. M. MARTÍNEZ-DUART

Catedrático de Física Aplicada
Universidad Autónoma de Madrid
Madrid, España

ADDISON-WESLEY/UNIVERSIDAD AUTÓNOMA DE MADRID

Argentina • Brasil • Chile • Colombia
Ecuador • España • Estados Unidos • México
Perú • Puerto Rico • Venezuela



CAPITULO I

SEMICONDUCTORES

Los semiconductores ocupan un lugar prominente en el conjunto de los materiales. Esto se debe al alto grado de desarrollo que se ha alcanzado en el conocimiento de sus propiedades básicas así como también en el de sus aplicaciones. Podemos decir que hoy día los semiconductores son piezas básicas en toda la tecnología electrónica, la cual en los últimos años ha mostrado un crecimiento espectacular, abarcando el campo de los procesadores, las comunicaciones, la robótica, etc. En este capítulo se pretende dar una descripción general del comportamiento de los semiconductores, y más en particular de las propiedades de conducción. Estas propiedades están determinadas fundamentalmente por la disposición de los electrones dentro de los átomos que forman el material semiconductor. De ahí surge la conveniencia de comprender los aspectos más básicos de la estructura electrónica de la materia y del enlace químico, que serán también tratados en este capítulo.

1.1. CLASIFICACION DE LOS MATERIALES DESDE EL PUNTO DE VISTA ELECTRICO.

Desde el punto de vista eléctrico, los materiales suelen dividirse en tres categorías atendiendo a su conductividad: conductores, semiconductores y aislantes. En la fig. 1.1 se han ordenado algunos materiales típicos según el valor de su conductividad. Nótese que la escala de conductividad tiene un rango muy amplio desde $10^{-18} \text{ ohm}^{-1} \text{ cm}^{-1}$ para los mejores aislantes hasta un valor mayor que $10^{26} \text{ ohm}^{-1} \text{ cm}^{-1}$ para los materiales superconductores a temperaturas por debajo de la temperatura crítica de transición. Los valores indicados en la fig. 1.1 han de tomarse como aproximados, ya que la conductividad es una magnitud sujeta a la influencia de muchos factores, tales como el estado de agregación del material, su estructura cristalina, temperatura, etc.

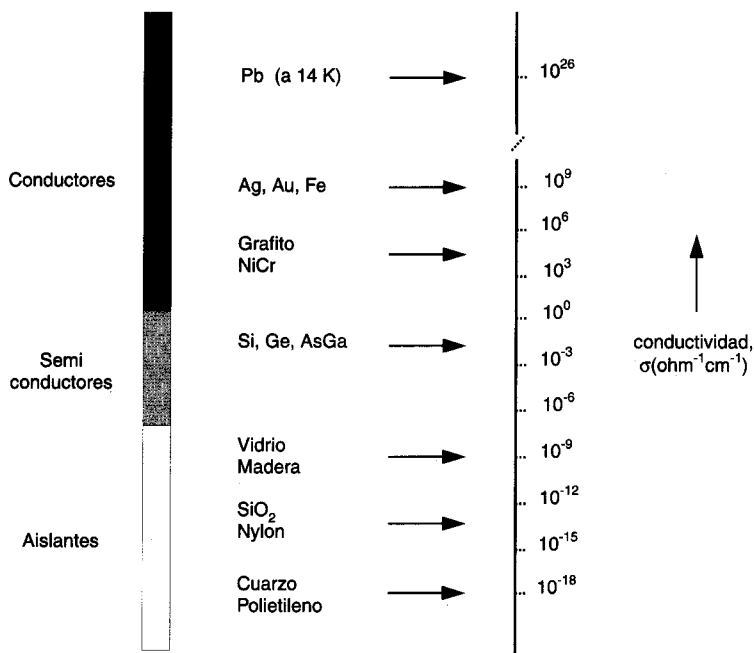


Fig. 1.1. Representación de los valores de la conductividad en algunos materiales típicos.

El primer dato a destacar en la fig. 1.1 es que la conductividad es una de las magnitudes físicas que admite mayor espectro de variación. Dejando aparte el caso de los materiales superconductores, la conductividad puede variar en más de 25 órdenes de magnitud al pasar de los materiales aislantes como el vidrio o los plásticos a los materiales conductores, tales como el cobre o la plata. Esto da lugar a que los diferentes tipos de materiales puedan presentar fenómenos eléctricos muy diversos. Así, en los materiales aislantes las propiedades eléctricas están dominadas por los llamados *fenómenos de polarización*, esto es, la deformación de la nube electrónica que rodea los átomos y las moléculas que componen el material por efecto del campo eléctrico aplicado, formando lo que se llama dipolos eléctricos. En los materiales conductores, por el contrario, los fenómenos predominantes al aplicar un campo eléctrico son los de conducción, debido al movimiento de electrones libres en el interior del material arrastrados por el campo eléctrico aplicado.

Los semiconductores forman un grupo de materiales que presenta un comportamiento intermedio entre los conductores y los aislantes. Como veremos más adelante, los semiconductores en estado puro y a temperaturas bajas presentan una conductividad relativamente baja por lo que sus propiedades se asemejan a las de los aislantes. Sin embargo, la conductividad

de estos materiales es una función creciente con la temperatura de forma que a la temperatura ambiente la mayoría de los semiconductores presentan una conductividad apreciable, aunque siempre es menor que la de un metal. Incluso a una temperatura dada, es posible variar a voluntad la conductividad de estos materiales si se les añade una cantidad controlada de impurezas de determinados elementos químicos. Es precisamente esta característica la que ha permitido desarrollar una gran variedad de componentes y dispositivos electrónicos basados en los materiales semiconductores.

1.2. ESTRUCTURA ELECTRONICA DE LOS MATERIALES SOLIDOS

Nos podemos preguntar pues, a qué obedecen las diferencias de comportamiento eléctrico entre unos materiales y otros. Como la mayor parte de las propiedades de estado sólido, estas diferencias están originadas por la diferente composición química y estructura electrónica de enlace de los átomos que forman el material. Así pues, es conveniente hacer un repaso de los aspectos más esenciales que determinan la estructura de enlace de los materiales.

Consideremos primero la estructura electrónica de **átomos aislados**, esto es sin interacción entre ellos. La mecánica cuántica nos dice que los electrones de los átomos se mueven alrededor del núcleo con una cierta energía que sólo puede tomar unos valores bien definidos (*orbitales o niveles atómicos*). El cálculo de la energía asociada a los niveles atómicos es generalmente complejo, y sólo es posible llevarlo a cabo de forma exacta para el átomo de hidrógeno, formado por un protón y un electrón. En este caso, considerando el núcleo en reposo, la energía del electrón ocupando un nivel n viene dada por la expresión:

$$E = - \frac{q^4 m_0}{8\epsilon_0 n^2 h^2} \quad [1.1]$$

donde q es la carga del electrón, m_0 su masa, ϵ_0 la permitividad del vacío y h la constante de Planck. Introduciendo los valores numéricos de estas magnitudes resulta para la energía del nivel n :

$$E = - \frac{13.6}{n^2} \text{ eV} \quad [1.2]$$

En el nivel más bajo ($n = 1$) la energía vale -13.6 eV. Este nivel se denomina *estado fundamental* del átomo de hidrógeno. El signo negativo indica que se trata de energía de enlace, esto es la energía necesaria para sacar el electrón desde el estado fundamental hasta

una posición fuera de la influencia del núcleo (a distancias infinitas). Esta energía, que para el hidrógeno es de 13.6 eV, se denomina también *energía de ionización*. El estado fundamental del electrón representa el estado de energía más baja. Sin embargo el electrón puede ocupar también otros estados de mayor energía (con $n > 1$), denominados *estados excitados*, cuando recibe energía suficiente mediante algún proceso de excitación (térmica, luminosa, etc.).

En la fig. 1.2a se da un esquema de los niveles energéticos del electrón para el átomo de hidrógeno, mostrando asimismo la curva de energía potencial, la cual sigue una ley del tipo: $E_{\text{pot}} = -q^2/4\pi\epsilon_0 r$, siendo r la distancia del electrón al centro del átomo. Esta curva determina la región espacial en la cual el electrón puede moverse bajo la influencia del núcleo, ya que fuera de los límites de la curva la energía total del electrón sería menor que la energía potencial y por tanto no hay estados posibles para el electrón. Los átomos con mayor número de electrones tienen también una estructura de niveles energéticos similar a la del hidrógeno, aunque están distribuidos y ordenados de manera más compleja, ya que el valor de la energía no depende solamente del valor de n sino también de otros números cuánticos. En cualquier caso, en el estado fundamental (es decir en el de energía más baja), los electrones del átomo se distribu-

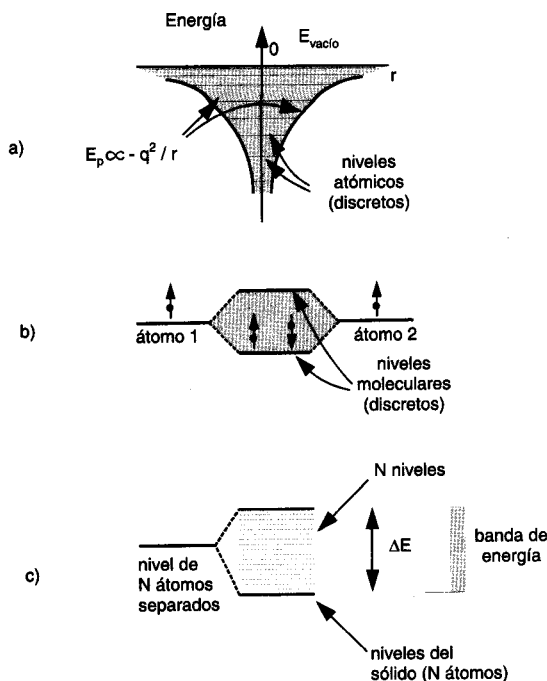


Fig. 1.2. Estructura de los niveles energéticos de los electrones: a) en un átomo aislado, b) en una molécula formada por dos átomos, y c) para un conjunto de N átomos. En los casos b) y c) sólo se muestra el desdoblamiento de los niveles correspondientes a los electrones de enlace.

yen entre los niveles más bajos de energía siguiendo el *principio de exclusión de Pauli*, el cual establece que en cada estado cuántico (definido por un conjunto de números cuánticos) no puede haber más de un electrón. Debido a la multiplicidad del espín, este principio implica que en cada nivel de energía sólo puede haber hasta un máximo de dos electrones.

Cuando se trata de **moléculas formadas por dos o más átomos**, es bien conocido que solamente los electrones de las capas más externas de cada átomo participan en el enlace interaccionando con el resto de los átomos, mientras que el resto de los electrones sigue unido a sus núcleos respectivos. En un enlace típicamente covalente, tal como el que presentan muchas moléculas diatómicas homopolares (H_2 , Cl_2 , etc.), la interacción de los electrones de enlace con el potencial eléctrico de los dos núcleos atómicos da lugar a un desdoblamiento de los niveles de energía originales de estos electrones. Resulta así en este caso dos niveles nuevos separados por una cierta energía, ΔE , uno de ellos con energía más baja que el nivel original (estado fundamental), y el otro con energía más elevada (estado excitado). Los electrones de enlace se sitúan en los nuevos niveles ocupando primero los de energía más baja. La ocupación se hace siguiendo también el principio de exclusión de Pauli, de forma que si por ejemplo cada uno de los átomos de la molécula aporta un electrón al enlace, los dos electrones se sitúan en el nivel inferior con los espines apareados, dejando el nivel superior vacío (susceptible de ser ocupado si la molécula se encuentra en un estado excitado) (fig. 1.2b). El estado fundamental se corresponde en este caso al de los electrones de enlace moviéndose alrededor de los núcleos de cada átomo, en los llamados *orbitales moleculares*, con una energía de movimiento bien definida. El resto de los electrones pertenecientes a las capas más internas de los átomos no participa en el enlace y por tanto se mantiene en los orbitales atómicos originales, con una energía que es prácticamente la misma que la que tenían con los átomos separados (no mostrados en la fig. 1.2b).

El caso de **moléculas aisladas con un cierto número de átomos** unidos entre sí mediante enlace covalente es mucho más complejo, aunque se pueden extrapolar algunas de las conclusiones del modelo descrito anteriormente para moléculas diatómicas. Así, se puede demostrar que en las moléculas simples formadas por N átomos iguales los electrones de las capas más internas se mantienen en sus niveles originales mientras que los electrones de enlace se sitúan en nuevos niveles originados por el desdoblamiento de los últimos niveles atómicos. Como resultado de este desdoblamiento aparecen N niveles nuevos (o subniveles) donde se sitúan los electrones de enlace, ocupando primero los de energía más baja. Hay que tener en cuenta que de acuerdo con el principio de exclusión de Pauli, cada uno de estos subniveles puede albergar hasta dos electrones, debido al apareamiento del espín. Un aspecto interesante a destacar en este caso es que la **diferencia de energía, ΔE , entre el subnivel más bajo y el más alto es prácticamente independiente del número total de átomos que forma la molécula**. Generalmente, esta diferencia de energía tiene un valor de unos pocos electrón-voltio, eV. (fig. 1.2c).

El fenómeno del desdoblamiento de los niveles atómicos es completamente general y se presenta también en los **sólidos con enlace covalente**. Este es el caso de la mayoría de los

semiconductores y también de los metales, en los cuales el número de átomos participantes en el enlace es muy elevado (alrededor de 10^{23} átomos por centímetro cúbico). En estos materiales, la interacción de los electrones de enlace con el conjunto de los N átomos del sólido da lugar al desdoblamiento de los niveles atómicos originales más elevados en un total de N nuevos subniveles. La diferencia de energía entre el subnivel máximo y el mínimo sigue siendo de unos pocos electrón-voltio, por lo que los subniveles individuales se encuentran muy próximos entre sí, es decir, separados por una energía extremadamente pequeña, ya que ahora el valor de N es muy elevado. La continuidad en energía de los subniveles permite hablar en este caso de una *banda de energía* con una anchura total de unos pocos electrón-voltio, constituida por N subniveles, y con capacidad de alojar hasta $2N$ electrones provenientes de los electrones de valencia de cada uno de los átomos. Si por ejemplo, cada átomo aporta un solo electrón la banda de energía queda ocupada hasta la mitad. En cambio, si aportan dos electrones por átomo la banda quedaría completamente ocupada. Esta banda de energía ocupada por los electrones de valencia se denomina *banda de valencia*. Es importante mencionar que los orbitales atómicos originales de los electrones de enlace forman por solapamiento de unos con otros nuevos orbitales que se extienden espacialmente por todo el sólido.



Fig.1.3. Estructura de los niveles energéticos y de las bandas de energía correspondientes a una red monodimensional de átomos.

Los niveles excitados de los átomos, incluso aunque no estén ocupados por electrones, también están sujetos a un desdoblamiento, dando lugar a la formación de bandas de energía cuando se trata de materiales sólidos con un número elevado de átomos. Así, en un sólido con enlace covalente los niveles excitados se convierten en una banda continua de niveles energéticos que se sitúa por encima de la banda de valencia, separada por una zona o "gap"¹ de energía donde no existen niveles. Se trata pues de una banda prohibida, en la que no existen estados energéticos posibles para los electrones de enlace.

¹ Nota: El término "gap", que procede del inglés, se utiliza para designar la banda de energía prohibida.

En la fig. 1.3 se da un esquema de los niveles energéticos de los electrones de un sólido ideal formado por una red de átomos monodimensional. En la figura se incluye la curva de energía potencial debida al conjunto de átomos, que limita también la extensión espacial del movimiento de los electrones. En los átomos ocupando posiciones extremas, esta curva se extiende hasta un valor de energía igual a cero, es decir, igual que en el caso de átomos aislados. A menudo a este nivel de energía se le conoce como *nivel de vacío* (señalado como $E_{\text{vacío}}$ en la fig. 1.3). Según se observa, entre cada átomo la curva de energía potencial alcanza un valor máximo por debajo del nivel de vacío, debido a la interacción de los átomos entre sí.

Hay que notar además que en los sólidos, al igual que en las moléculas aisladas, los electrones de las capas más internas se sitúan en niveles discretos limitados espacialmente por la curva de energía potencial de cada átomo, mientras que los de las capas más externas que participan en el enlace se distribuyen entre los niveles que forman la banda de valencia. Por encima de esta banda se encuentra la banda correspondiente a los estados excitados, también denominada *banda de conducción* por razones que se harán aparentes más adelante. Nótese que tanto la banda de valencia como la de conducción no están limitadas espacialmente y por tanto se extienden por toda la cadena de átomos. Esto significa que en un sólido con enlace covalente los electrones de enlace tienen cierta capacidad de movimiento a través de todo el cristal intercambiando su posición unos con otros. A temperaturas próximas al cero absoluto todos los electrones de enlace ocupan los niveles más bajos de la banda de valencia. Sin embargo a temperaturas superiores, algunos electrones pueden ser excitados a otros niveles más elevados dentro de esta banda, siempre que existan niveles o estados vacantes. Incluso pueden pasar a la banda superior de energía si los electrones adquieren energía suficiente para saltar el “gap”.

Son los electrones pertenecientes a estas bandas de energía (bandas de valencia y de conducción) los que confieren las propiedades eléctricas características de los materiales. La energía de estas bandas, su separación, el número de electrones en cada banda, etc., está determinado en parte por factores intrínsecos del material tales como el tipo de enlace, distancia entre los átomos, etc. y también por factores extrínsecos al material, como la temperatura, contenido de impurezas, etc. En lo que sigue, centraremos nuestro estudio en el comportamiento eléctrico de los materiales según el modelo de bandas de energía.

1.3. CONDUCTORES, SEMICONDUCTORES Y AISLANTES

En la fig. 1.4 se presenta un esquema de la estructura de bandas de energía típica de los materiales conductores, semiconductores y aislantes a temperaturas próximas a 0 K. En este esquema, el eje horizontal representa una de las coordenadas espaciales dentro de la red de átomos, mientras que el eje vertical indica la energía total de los electrones dentro de cada banda. En este tipo de diagramas normalmente se prescinde por comodidad del nivel de vacío, ya que a menudo lo que interesa es conocer la diferencia de energía del electrón en

relación a los bordes superior e inferior de la banda de valencia y de conducción, respectivamente. Según se ha mencionado anteriormente, las dos bandas representadas son las últimas que pueden estar ocupadas por los electrones de enlace del material y reciben el nombre de banda de valencia, la de energía inferior, y banda de conducción la de energía más elevada. Ambas bandas están separadas por una zona de energía prohibida, de valor E_g , denominada *banda prohibida*.

En las proximidades del cero absoluto los electrones de la banda de valencia ocupan los niveles más bajos de energía, formando parte del enlace entre los átomos (zona rayada en la fig. 1.4). En el caso de que la banda no se encuentre totalmente ocupada, los niveles más altos de energía dentro de la banda permanecerán vacíos. En cambio, a temperaturas más elevadas, una fracción apreciable de los electrones puede ser excitada a niveles con energía superior, siempre que estos niveles (o posiciones de enlace) se encuentren vacantes. Esta circunstancia, es decir, la existencia de niveles vacantes en una banda de energía, es la que permite que los electrones se puedan mover dentro del cristal bajo la acción de un campo eléctrico. Esta condición viene impuesta por las leyes de la mecánica cuántica, ya que el movimiento de los electrones implica una ganancia en energía cinética y por tanto en su energía total. Este aumento de energía sólo se produce si existen niveles de energía vacantes en la banda de valencia, de forma que los electrones en su movimiento puedan pasar a ellos. El movimiento de los electrones en el interior del cristal es bastante complejo, ya que no sólo actúan las fuerzas del campo eléctrico aplicado sino también las debidas a la interacción de los electrones con los átomos del cristal. Más adelante, en el capítulo siguiente, analizaremos con más detalle los procesos microscópicos que tienen lugar durante la conducción.

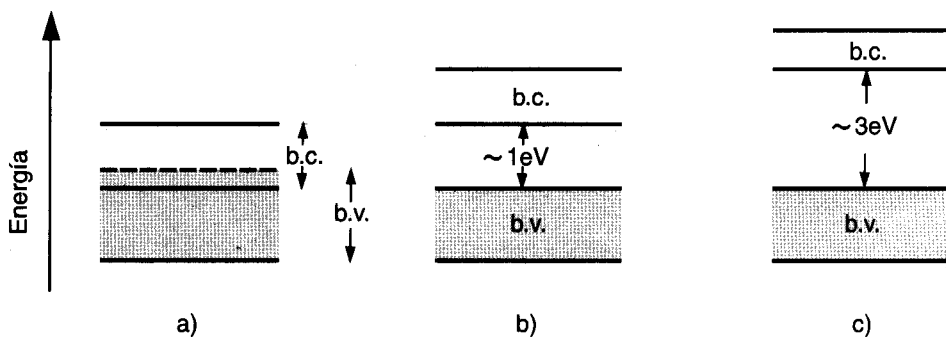


Fig. 1.4. Estructura de bandas en materiales de tipo: a) conductor, b) semiconductor, y c) aislante, a temperaturas próximas al cero absoluto.

En el caso opuesto, si la banda de energía se encuentra totalmente ocupada, la aplicación de un campo eléctrico no implica necesariamente un desplazamiento neto de electrones en la dirección del campo, aún cuando los electrones pueden tener una cierta movilidad dentro de la banda. La ausencia de niveles vacantes da lugar a que el desplazamiento de un electrón en la dirección del campo esté compensado siempre por el de otro electrón en sentido opuesto. De todo esto se deduce que la **presencia de electrones en una banda, de valencia o de conducción, en la cual existen niveles o estados vacantes a los cuales el electrón se pueda trasladar, es una de las condiciones que se exige para que los electrones de esa banda puedan participar en los procesos de conducción al aplicar un campo eléctrico.**

Para comprender mejor este principio se puede recurrir a la analogía mecánica de la fig. 1.5, en la cual se representa un tubo de vidrio herméticamente cerrado y con agua en su interior. Cuando el tubo está completamente lleno (fig. 1.5a), no se observa movimiento del líquido incluso cuando se inclina el tubo. En cambio cuando el tubo no está completamente lleno, bien sea debido a que el líquido ocupa sólo un cierto volumen del tubo (fig.1.5b), o bien debido a la presencia de burbujas de aire (fig.1.5c), el agua se pone inmediatamente en movimiento al inclinar el tubo. En el caso de las burbujas, el movimiento se detecta por el desplazamiento del aire en sentido opuesto al del líquido. Es esta circunstancia la que pone de manifiesto la necesidad de que exista al menos un pequeño volumen desocupado para que se pueda producir el movimiento del agua.

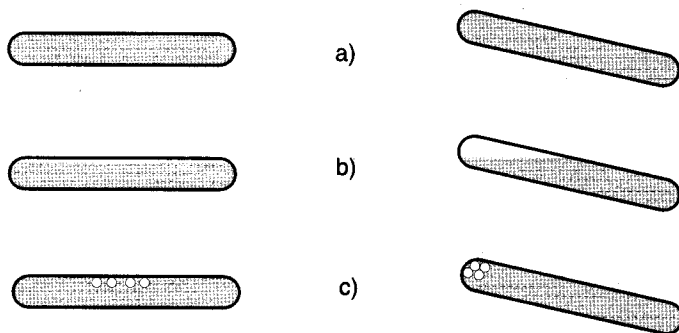


Fig.1.5. Analogía mecánica del fenómeno de la conducción debido al desplazamiento de los electrones a niveles vacantes de energía. a) Si el tubo está lleno de agua no se observa desplazamiento del líquido al inclinarlo. En cambio, cuando el tubo está parcialmente lleno, b), o contiene burbujas, c), sí se observa movimiento del líquido.

A partir de estos hechos es fácil explicar el comportamiento eléctrico de los diferentes tipos de materiales. En el caso de los **metales**, existe un enlace covalente “compartido” entre todos los átomos, en el que cada átomo aporta uno o varios electrones de enlace, según sea la valencia del metal. La estructura de bandas de los metales generalmente presenta una situación peculiar, ya que la banda de conducción solapa en energía con la banda de valencia de forma que no existe banda de energía prohibida (fig. 1.4a). Los electrones se encuentran por tanto dentro de una banda única de energía, que a menudo es referida como la banda de conducción del metal, en la cual, lógicamente, existen numerosos niveles vacantes. En el cero absoluto todos los electrones de esta banda se hallan ocupando los niveles energéticos más bajos hasta un cierto valor de la energía, y por encima de este valor el resto de los niveles se encuentra vacante. A temperaturas superiores al cero absoluto, la existencia de niveles vacíos con energías más elevadas hace que los electrones se puedan trasladar a ellos mediante algún proceso de excitación térmica. Cuando se aplica un campo eléctrico los electrones se desplazan saltando entre niveles vacantes, participando así en la conducción. En los metales, el número de electrones presentes en la banda de conducción es muy elevado. Así, en los metales monovalentes existe en esta banda alrededor de un electrón por cada átomo del material, es decir, del orden de 10^{22} ó 10^{23} electrones por cm^3 , mientras que el número de niveles permitido es el doble. Es evidente, pues, que la conductividad en estos materiales ha de ser muy elevada.

En los **semiconductores** con enlace típicamente covalente, cada átomo aporta también un número determinado de electrones para formar el enlace con los átomos vecinos. A temperaturas próximas a las del cero absoluto todos los electrones de valencia participan en el enlace de unos átomos con otros y la banda de valencia se halla completamente llena, es decir, sin estados vacantes, mientras que la de conducción está completamente vacía, por lo que en estas condiciones no puede haber conducción. Sin embargo, la energía de enlace de los electrones es relativamente pequeña de forma que a temperaturas ordinarias (300 K) una fracción apreciable de electrones puede romper el enlace y pasar a la banda de conducción donde existe un gran número de estados vacantes. Estas vacantes, junto con las generadas en la banda de valencia, hacen que los electrones puedan participar en los procesos de conducción cuando se aplica un campo eléctrico. La energía necesaria para romper el enlace se corresponde con la energía de la banda prohibida, con un valor alrededor de 1 eV, o incluso menor, para la mayoría de los semiconductores (fig. 1.4b). En un semiconductor típico como el silicio, el número de electrones que pueden pasar a la banda de conducción a la temperatura ambiente es del orden de 10^{10} electrones/ cm^3 . Por ello su conductividad, aunque apreciable, será mucho más baja que la de los metales. Podemos decir que los materiales semiconductores a temperaturas bajas tienen un comportamiento típico de los materiales aislantes, descritos más abajo, mientras que a temperaturas medias o altas su comportamiento se acerca más al de los metales, al poseer un cierto número de electrones disponibles para la conducción.

En los materiales **aislantes** con enlace covalente, fig. 1.4c, los electrones de enlace están compartidos por cada pareja de átomos, formando un enlace muy fuerte y ocupando completamente los estados de la banda de valencia. Del mismo modo, en los materiales con

enlace iónico, los electrones de valencia se encuentran también muy unidos a los iones respectivos formando una banda muy estrecha de energía. Todos estos materiales requieren una energía bastante elevada para romper el enlace de forma que los electrones puedan pasar a la banda de conducción. En consecuencia, los aislantes tienen la banda de conducción separada en varios eV de la banda de valencia, por lo que a temperaturas ordinarias todos los niveles de la banda de valencia están ocupados, mientras que los de la banda de conducción se encuentran prácticamente vacíos de electrones. Todo ello hace que a la temperatura ambiente la conductividad de los aislantes sea muy baja, ya que la banda de conducción está vacía de electrones y, por otra parte, no existe posibilidad de que los electrones de la banda de valencia puedan ser arrastrados por un campo eléctrico externo. Sin embargo, en algunos materiales aislantes se observa una cierta conductividad por encima de los valores esperados. En este caso la conducción se debe a factores externos, tales como la presencia de impurezas u otros defectos del material. Estos agentes dan lugar a una cierta inyección o trasvase de electrones a la banda de conducción, aumentando así de forma sensible las propiedades conductoras del material.

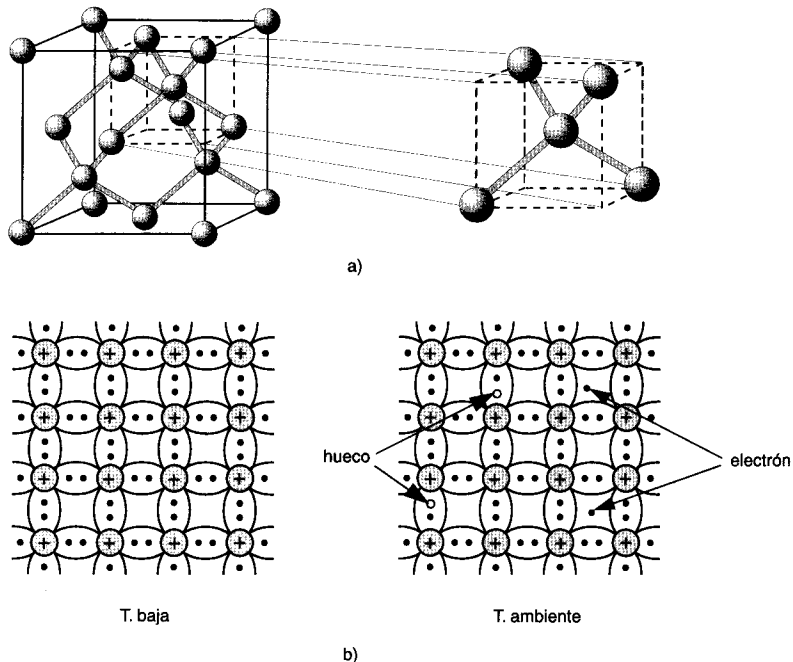


Fig.1.6. a) Estructura de enlace del silicio, similar a la red del diamante. b) Representación bidimensional de la estructura de enlace del silicio. Fig. izquierda: a bajas temperaturas mostrando todos los enlaces saturados. Fig. derecha: a temperatura ambiente, con algunos enlaces sin saturar dando lugar a huecos y electrones libres.

1.4. SEMICONDUCTORES INTRINSECOS

Veamos con más detalle las propiedades conductoras de los llamados semiconductores *intrínsecos*, es decir, la de aquellos semiconductores en estado puro y perfectamente cristalizados. El silicio es uno de los elementos semiconductores típicos empleados en la fabricación de la mayor parte de los dispositivos electrónicos de estado sólido. En estado cristalino, los átomos de este material ocupan posiciones tetraédricas en una red similar a la del diamante, compartiendo cuatro electrones con sus átomos vecinos en un enlace de tipo covalente según se indica en la fig. 1.6a.

En la fig. 1.6b se da una representación bidimensional de la estructura y del enlace químico de un semiconductor típico, de valencia 4, como el silicio o el germanio, en los cuales cada átomo comparte dos electrones con otro átomo vecino. Otros compuestos, también con características semiconductoras, están formados por la combinación de elementos del grupo III y el grupo V del sistema periódico, como el arseniuro de galio, GaAs, fosfuro de galio, GaP, etc., o bien por la combinación de elementos del grupo II y del grupo VI, tales como el sulfuro de zinc, ZnS, o el telururo de cadmio, CdTe. En todos estos casos, tanto la estructura cristalina del compuesto como su estructura de enlace es muy similar a la del silicio o germanio.

A temperaturas bajas, próximas al cero absoluto, todos los enlaces de los átomos se encuentran saturados con los electrones correspondientes, por lo que la banda de valencia se encuentra totalmente ocupada. Sin embargo, la energía necesaria para romper el enlace es relativamente pequeña, del orden de 1.1 eV para el silicio y 0.7 eV para el germanio. Los electrones pueden recibir esta energía mediante excitación térmica, por ejemplo. Los procesos de excitación térmica ocurren cuando se eleva la temperatura del material. Algunos electrones entonces ganan energía a partir de las vibraciones de los átomos, en cantidad suficiente para romper el enlace, pasando desde la banda de valencia a la de conducción, donde existen numerosos estados o niveles de energía vacantes (fig. 1.6b).

1.4.1. Portadores de carga: concepto de hueco

Según hemos visto, la excitación de un electrón a la banda de conducción implica la ruptura de un enlace en algún punto del cristal, donde a su vez se origina un *estado vacante* que además presenta una deficiencia de carga negativa (equivalente a una carga positiva de magnitud igual a la carga del electrón). Esta deficiencia de carga asociada a la vacante posee una cierta movilidad en el interior del cristal. De manera gráfica, la movilidad de los huecos se explica si se tiene en cuenta que los electrones que se encuentran en enlaces próximos a la vacante pueden saltar a esa posición vacante dejando tras sí una nueva vacante o enlace sin saturar. Este proceso da lugar a un desplazamiento de la vacante en sentido opuesto al del

electrón que efectúa el salto. Mediante un intercambio repetido de la vacante con los electrones de enlace próximos se origina un movimiento de la vacante de un punto a otro del cristal con un consumo de energía muy pequeño. Es más, la deficiencia de carga negativa asociada a un nivel vacante en la banda de valencia mantiene su entidad una vez que ha sido creada. Debido a ello, los niveles vacantes de la banda de valencia tienen un comportamiento muy similar al de los electrones de la banda de conducción. En realidad, el desplazamiento de estas vacantes electrónicas se hace mediante un proceso mas complejo, de naturaleza cuántica. Sin embargo, el modelo anterior puede ser suficiente para entender el comportamiento de los niveles vacantes de energía en la banda de valencia.

Estas características de los niveles o estados vacantes, denominados también *huecos*, permite considerarlos como partículas inmersas en un mar de electrones de enlace dentro de la banda de valencia, es decir, en proporción mucho menor que la de los electrones en esta banda. A la temperatura ambiente, solamente un electrón de cada 10^{12} de la banda de valencia, en el caso del silicio, rompe su enlace por excitación térmica para pasar a la banda de conducción, dejando tras sí el correspondiente nivel vacante. Además, debido a que los huecos son capaces de moverse en el interior del cristal también se les puede asociar una energía cinética de movimiento (en realidad se trata de la energía de los electrones que se desplazan en sentido

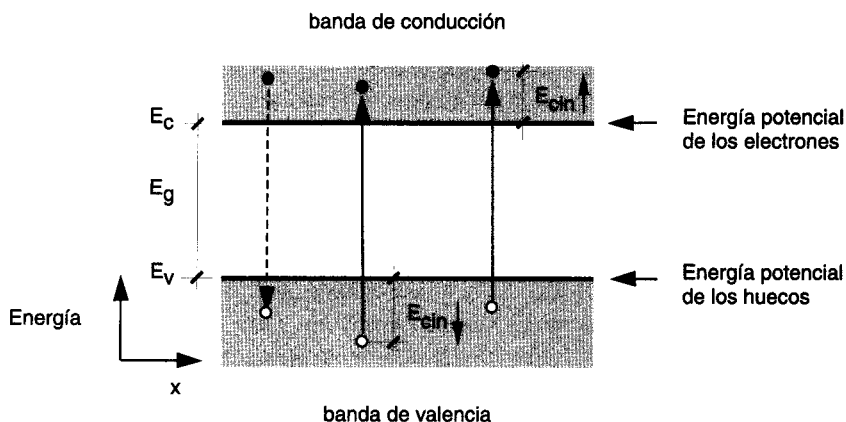


Fig.1.7. Diagrama de bandas de energía de un semiconductor mostrando las componentes de energía cinética y energía potencial de los electrones y huecos (en la figura se ha representado sólo los estados próximos al tope de la banda de valencia y al fondo de la banda de conducción).

opuesto ocupando las posiciones del hueco). A partir de estos hechos se desprende la importancia del hueco como entidad, con un comportamiento similar al de los electrones que se

mueven en la banda de conducción. Tal es así que, desde el punto de vista cuantitativo, el hueco es también considerado como una partícula que se mueve con una energía dentro de la banda de valencia y que posee una carga igual a la del electrón pero de signo opuesto, es decir positiva. Debido a ello los huecos pueden ser arrastrados con una cierta movilidad por un campo eléctrico externo, según veremos más abajo. Incluso es posible también asociar a los huecos una cierta masa, denominada *masa efectiva*. Estas características hacen que tanto los huecos como los electrones de un semiconductor intrínseco sean denominados indistintamente *portadores de carga o portadores intrínsecos*.

1.4.2. Interpretación del esquema de bandas de energía

En el esquema de bandas de energías, la escala vertical representa la **energía total**, E , de los niveles electrónicos en la banda de valencia o de conducción. En este esquema, el valor de E_g corresponde a la energía mínima necesaria para romper un enlace en el cristal, lo que a su vez implica que tanto el electrón como el hueco generado en el proceso quedan en reposo en los niveles de energía correspondientes al fondo de la banda de conducción, E_c , y al tope de la banda de valencia, E_v , respectivamente. En este proceso de ruptura de un enlace, cualquier exceso de energía absorbida sobre el valor de E_g obedece bien sea a que el electrón liberado procede de un nivel inferior a E_v en la banda de valencia, o incluso a que el nivel de destino en la banda de conducción posee una energía mayor que E_c . En uno u otro caso, se imparte al electrón o al hueco creado una energía adicional que se traduce en definitiva en energía de movimiento a través del cristal. Así por ejemplo, si E representa la energía final del electrón una vez que pasa a la banda de conducción, la diferencia $E - E_c$ se interpreta como la **energía cinética** asociada al movimiento del electrón en la banda de conducción. Del mismo modo, si el hueco generado en la banda de valencia se encuentra en un nivel de energía E (por supuesto diferente a la del electrón), la diferencia $E_v - E$ representa la energía cinética del hueco. En uno y otro caso, puesto que E representa la energía total de la partícula, esto es la suma de las energías cinética y potencial, $E = E_{cin} + E_{pot}$, los valores de E_c y E_v representan a su vez la **energía potencial** (asociada al enlace) medida desde un cierto nivel de referencia para los electrones en la banda de conducción y los huecos en la banda de valencia, respectivamente (fig. 1.7).

Así pues, en el diagrama de bandas de energía, la energía cinética de los electrones de la banda de conducción viene dada por el valor de E medido desde E_c . Por el contrario, para los huecos la energía cinética viene dada por el valor de E medido desde E_v hacia abajo. Dado que E_c y E_v corresponden al valor de la energía potencial, se puede tomar para esos niveles un origen arbitrario. Por esta razón, es muy frecuente omitir el nivel de referencia en un esquema de bandas, aunque a veces en los casos que se hace necesario se toma como referencia el nivel de vacío o infinito, mencionado más arriba. Este nivel se halla generalmente situado varios electrón-voltio por encima del fondo de la banda de conducción.

Las leyes de la mecánica cuántica predicen para un electrón que se mueve en la banda de conducción sometido al potencial periódico de los átomos una relación entre su energía cinética, $E_{\text{cin}} = E - E_c$, y el momento cuántico, p , similar a la que existe para un electrón libre ($E_{\text{cin}} = p^2/2m_0$, con m_0 = masa del electrón libre). Para un electrón con energía E dentro de la banda de conducción tendremos:

$$E = E_c + p^2/2m_e^* \quad [1.3a]$$

siendo p , el llamado momento cristalino, esto es el equivalente cuántico del vector momento y m_e^* la masa efectiva del electrón mencionada más arriba. Debido a que el electrón durante su movimiento, está interaccionando constantemente con los átomos de la red cristalina, la masa efectiva no es una constante en el sentido estricto, sino que depende de factores tales como la estructura cristalina, la disposición de los átomos en una dirección determinada, etc. Por esta razón, m_e^* está influenciada incluso por la dirección del movimiento del electrón. Algo similar ocurre para los huecos, para los cuales se puede considerar que poseen una cierta masa efectiva, m_h^* , determinada por una ecuación análoga a la anterior, es decir:

$$E = E_v - p^2/2m_h^* \quad [1.3b]$$

En el caso del silicio, por ejemplo, la masa efectiva de los electrones moviéndose en la dirección $[100]$ tiene un valor dado por: $m_e^* = 0.19m_0$. Generalmente, la masa efectiva de los huecos suele ser algo mayor que la de los electrones. El que las masas efectivas de electrones y huecos sean diferentes no debe sorprendernos, si se considera que el movimiento de los electrones en la banda de conducción es relativamente libre (aún cuando interaccionan con los átomos de la red), en cambio el de los huecos implica un intercambio de enlaces entre los átomos de la red.

1.4.3. Fenómenos de conducción

De la discusión precedente se desprende la posibilidad de que tanto los electrones en la banda de conducción como los huecos en la de valencia puedan participar directamente en los procesos de conducción. Efectivamente, supongamos que aplicamos sobre una barra semiconductor uniforme de longitud L una diferencia de potencial $V = V_2 - V_1$, con $V_2 > V_1$. Este potencial eléctrico se superpone al potencial al que están sometidos los electrones en el interior del cristal y hace que su energía potencial varíe en la cantidad $-qV_2$ para los electrones que se encuentran en un extremo del semiconductor y en $-qV_1$ para los que se encuentran en el extremo contrario. Si para mayor simplicidad hacemos $V_1 = 0$, podemos decir que la energía potencial de los electrones disminuye uniformemente desde uno de los lados hasta el otro,

hasta alcanzar el valor $-qV_2 = -qV$. En un esquema de bandas de energía, esto implica que las líneas que representan los valores de E_c y E_v deben representarse inclinadas, según se indica en la fig. 1.8a, con una caída total igual a $q\Delta V$. Además, si el semiconductor es uniforme, el campo eléctrico en su interior asociado al potencial aplicado debe ser constante, con un valor E dado por $E = -\Delta V/L$. Esto quiere decir que la pendiente de las bandas de energía representadas en la fig 1.8a coincide en valor absoluto con el valor del campo eléctrico. A lo largo de este tratado encontraremos numerosos ejemplos de inclinación de las bandas de energía por efecto de la presencia de un campo eléctrico actuando en el semiconductor. Conviene recordar que en todos estos casos **la pendiente de las bandas energía en cada punto del semiconductor es proporcional al valor del campo eléctrico en ese punto, cambiado de signo.**

Como sabemos, el movimiento de los electrones se verifica en dirección opuesta al campo, desplazándose siempre hacia los puntos de energía potencial más baja. Así pues, la

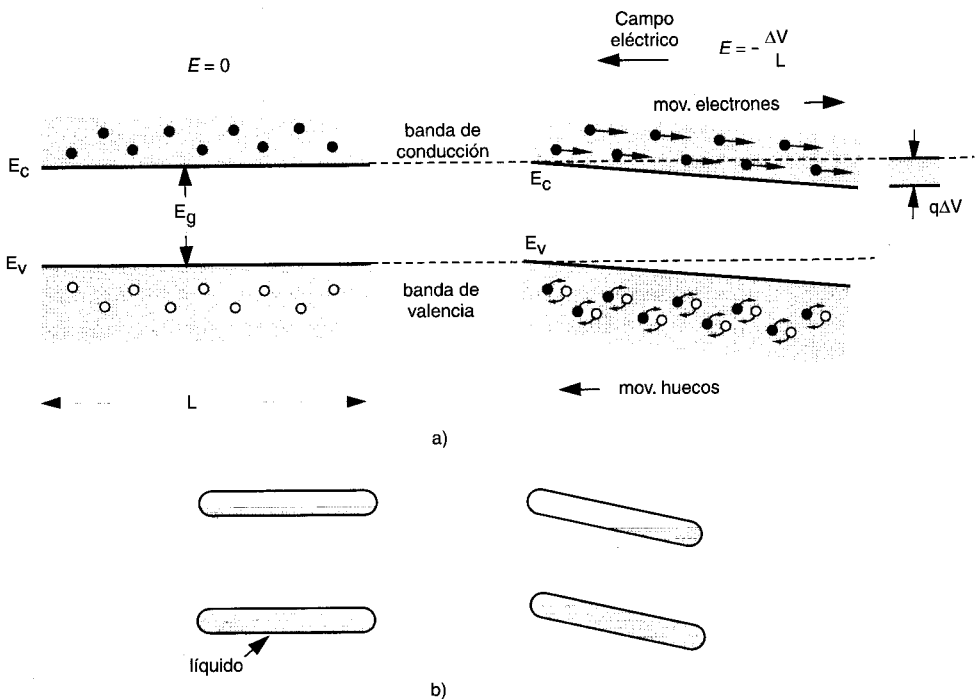


Fig.1.8. a) Esquema de bandas de energía de un semiconductor antes y después de aplicar un campo eléctrico. b) Analogía mecánica del movimiento de los electrones en la banda de conducción (tubo superior) y de los huecos en la banda de valencia (tubo inferior).

aplicación del campo eléctrico hace que los electrones de la banda de conducción se muevan dentro de la banda bajando hacia los puntos de menor potencial. Igualmente, cuando se trata de la banda de valencia también puede existir movimiento de electrones siempre que exista un hueco o estado vacante en sus proximidades, según hemos visto más arriba. Cuando esto ocurre, el electrón viaja del mismo modo hacia los puntos de energía potencial más baja, intercambiando su posición con el hueco correspondiente, lo cual se traduce en definitiva en un desplazamiento del hueco en sentido contrario al del electrón. De todo esto se concluye que los huecos de la banda de valencia se desplazan en la dirección del campo eléctrico o, lo que es lo mismo, hacia valores de energía potencial más elevada.

Para entender mejor el movimiento de los huecos por la acción de un campo eléctrico, conviene traer de nuevo la analogía mecánica del tubo de agua ya vista anteriormente. En la fig. 1.8b se han representado dos tubos, de los cuales al inferior se le ha extraído una pequeña cantidad de agua, dejando unas burbujas de aire, y esta cantidad se ha pasado al tubo superior. Es evidente que al inclinar los tubos (fuerza externa) las burbujas de aire del tubo inferior se moverán hacia arriba mientras que la pequeña cantidad de agua del tubo superior se desplazará hacia abajo. De una manera esquemática y sin pretender llevar la analogía hasta el fin, se puede describir el comportamiento de los electrones en la banda de conducción por acción del campo eléctrico como el de partículas clásicas que se desplazan hacia posiciones de energía potencial menor, mientras que el de los huecos es equivalente al de burbujas que tienden a “flotar” dentro de un mar de electrones, desplazándose hacia posiciones de energía potencial más elevada.

1.4.4. Fenómenos de excitación de portadores

En un semiconductor que se encuentra en equilibrio térmico a una temperatura dada, existe un proceso continuo de excitación de electrones desde la banda de valencia a la de conducción. En este proceso se rompe un enlace y se crea el hueco correspondiente en la banda de valencia. Los procesos de excitación están a su vez compensados por procesos de recombinación que actúan en sentido opuesto, en los cuales un electrón de la banda de conducción se desexcita y pasa a ocupar un nivel vacante de la banda de valencia, con lo que desaparece un hueco. De todo esto se desprende que **en un semiconductor intrínseco, en equilibrio térmico, la concentración de electrones presentes en la banda de conducción, n , debe ser igual a la de huecos en la banda de valencia, p , es decir: $n = p$.** Es más, el valor de n y p debe ser constante con el tiempo si la temperatura del material es constante, a pesar de que continuamente existen procesos de generación y de recombinación de pares electrón-hueco.

En los procesos de excitación térmica los electrones de enlace ganan energía de la red como consecuencia de las vibraciones de los átomos. Estas vibraciones dan lugar a la ruptura de un cierto número de enlaces produciéndose electrones “libres” en la banda de conducción y el correspondiente número de huecos en la banda de valencia. Si llamamos n_i a la concentra-

ción de electrones (o huecos) en la banda de conducción (valencia), el valor de n_i debe ser más elevado cuanto mayor sea la temperatura del cristal, ya que en este caso la energía de las vibraciones de la red es mayor. Así pues, para un semiconductor intrínseco que se encuentra en equilibrio térmico a una cierta temperatura T podemos escribir:

$$n = p = n_i(T) \quad [1.4]$$

El valor de n_i también depende, obviamente, del valor de la energía de la banda prohibida, E_g , ya que cuanto menor sea E_g mayor es el número de electrones que tiene energía suficiente de excitación para pasar desde la banda de valencia a la de conducción a una temperatura dada.

Un hecho a destacar en los semiconductores es que los electrones excitados a la banda de conducción y los huecos generados en la banda de valencia no se encuentran en una posición estática sino que están en continuo movimiento en el interior del cristal. La energía de movimiento procede también de la energía térmica del cristal, es decir, la impartida por las vibraciones de los átomos de la red del semiconductor.

En algunos semiconductores como el arseniuro de galio, denominado de "gap" directo, para que se efectúe la transición de una banda a otra sólo se requiere que la energía transferida al electrón sea igual o mayor que la energía de la banda prohibida. Por contra, los semiconductores de "gap" indirecto -el silicio y el germanio son los ejemplos más representativos- requieren por consideraciones de conservación del momento no sólo el aporte necesario en energía sino también la transferencia de una cierta cantidad de movimiento al electrón. Como veremos en el capítulo quinto, esta distinción entre semiconductores de "gap" directo e indirecto es muy importante en dispositivos optoelectrónicos, ya que la absorción o emisión de luz que ocurre en los procesos de excitación o de desexcitación respectivos se produce con la participación de un fotón, cuyo momento es muy pequeño. Por esta razón los procesos de emisión o absorción son más eficientes en semiconductores de "gap" directo.

1.5. SEMICONDUCTORES EXTRINSECOS

Según acabamos de ver, en un semiconductor intrínseco, la concentración de portadores intrínsecos -electrones o huecos- está determinada por la temperatura del semiconductor, $n_i = n_i(T)$, siendo $n_i(T)$ una función creciente con la temperatura. Más adelante veremos también que la función $n_i(T)$ tiene una naturaleza cuasi-exponencial. Así pues, la conductividad de un semiconductor puro es una magnitud muy sensible a los cambios de temperatura. Desde un punto de vista práctico, la variación de la conductividad con la temperatura constituye un

serio inconveniente para la utilización de los semiconductores en estado puro en la fabricación de dispositivos electrónicos, ya que en este caso lo que se pretende es que los dispositivos tengan un comportamiento lo más estable posible con la temperatura. Aparte de ello, interesa también disponer de semiconductores en los que la conducción esté determinada por un solo tipo de portadores, bien sea electrones o huecos.

Existe un procedimiento para obtener un valor relativamente constante del número de portadores de un semiconductor a la temperatura ambiente. Este procedimiento consiste en la introducción de átomos de diferente valencia, en una proporción adecuada, dentro de la red del material semiconductor. En los semiconductores de valencia +4, como el Ge o el Si, los átomos añadidos suelen ser de valencia +3, boro o aluminio, por ejemplo, o de valencia +5, fósforo, arsénico, antimonio, etc. La conductividad pasa entonces a estar dominada por la concentración y la naturaleza de los átomos añadidos, también denominados impurezas, y el semiconductor se denomina *extrínseco o dopado*.

1.5.1 Semiconductor de tipo n

Veamos cómo afecta la presencia de estas impurezas a la concentración de portadores, electrones y huecos, del material semiconductor. En la fig. 1.9a se ha dibujado de nuevo la

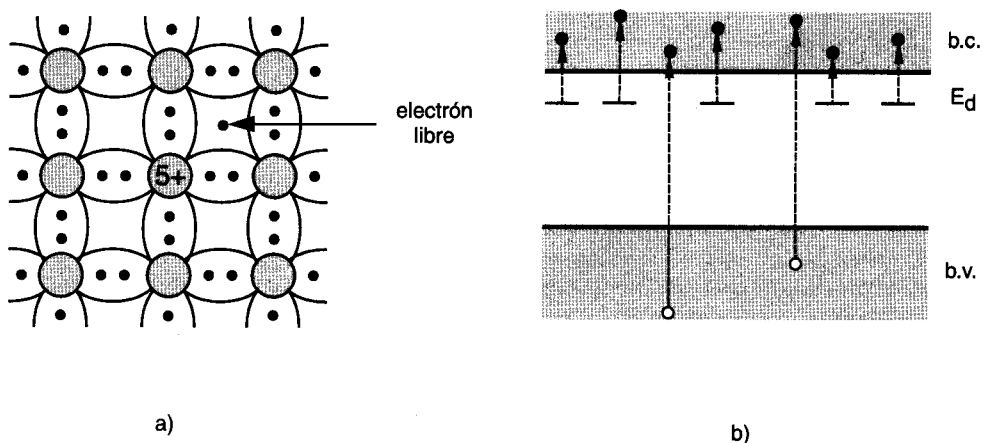


Fig. 1.9. a) Esquema de la estructura de enlace químico del silicio dopado con átomos pentavalentes. b) Esquema correspondiente de la estructura de bandas de energía de los electrones.

estructura de enlace de un semiconductor como el silicio, de valencia +4, dopado con una cierta proporción de átomos de antimonio. La valencia de este elemento es +5. Cuando la red del semiconductor está libre de defectos y formada por un cristal único o monocristal, los átomos de antimonio ocupan posiciones en sustitución de átomos de silicio cediendo también cuatro electrones para compartir con sus átomos vecinos de silicio en un enlace covalente. Cada átomo de antimonio tiene además un quinto electrón que no participa en el enlace y que queda por tanto muy débilmente ligado al propio átomo. Por esta razón a temperaturas próximas a la del ambiente este electrón recibe suficiente energía térmica para romper su enlace con el antimonio quedando en completa libertad para moverse a través del cristal. El átomo de impureza queda entonces ionizado con una carga positiva. Desde un punto de vista energético esta situación se representa mediante el paso del electrón desde un cierto nivel de energía (E_d), correspondiente a la energía de enlace con el átomo de antimonio, a la banda de conducción, con E_d situado en el interior del “gap” a unas centésimas de electrón-voltio de energía por debajo de E_c (fig. 1.9b).

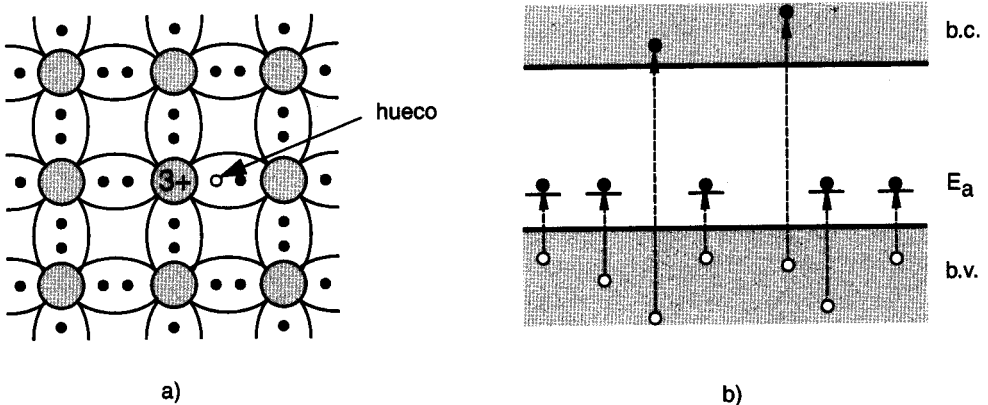


Fig. 1.10. a) Esquema de la estructura de enlace químico del silicio dopado con átomos trivalentes. b) Esquema correspondiente de la estructura de bandas de energía de los electrones.

A este tipo de impurezas, que como el antimonio son capaces de ceder electrones a la banda de conducción, se las denomina *donadoras*. Los electrones así generados, denominados “extrínsecos”, una vez en la banda de conducción son completamente indistinguibles de los electrones “intrínsecos”, y por tanto participan en pie de igualdad en los procesos de conducción. Observemos en la fig. 1.9b que la concentración total de electrones, n , en el

semiconductor extrínseco puede ser completamente diferente a la de huecos, p , a diferencia de los semiconductores intrínsecos en los que siempre se cumple que $n = p$ (ec. 1.4). La introducción de impurezas de tipo donador en cantidad suficiente hace que a la temperatura ambiente sea $n > p$, y el semiconductor se dice entonces que es de *tipo n* .

1.5.2. Semiconductor de tipo p

Si en lugar de impurezas de valencia +5, se añaden átomos de valencia +3, como el boro, ocurre una situación en ciertos aspectos similar a la anterior (fig. 1.10a). Los átomos ocupan también en este caso, por consideraciones de tamaño, posiciones sustitucionales, y comparten los tres electrones de la capa más externa con el resto de los átomos vecinos de silicio. La deficiencia de un electrón para completar el enlace hace que electrones vecinos que también participan en el enlace puedan ocupar esa vacante si son activados mediante energía. La energía necesaria para la activación suele ser muy pequeña de forma que a la temperatura ambiente prácticamente todos los átomos de impureza están ionizados, es decir han recibido un electrón “extra” para completar el enlace. Mediante este proceso se origina igual número de vacantes de electrones de enlace que átomos de impureza añadidos. El comportamiento eléctrico de estas vacantes es, para todos los efectos, igual al de los huecos generados térmicamente en la banda de valencia (fig. 1.10b). Así pues, mediante la adición de impurezas de valencia inferior, *impurezas aceptadoras o aceptoras*, aparece en la banda de valencia a la temperatura ambiente un número de huecos aproximadamente igual al de átomos de impurezas debido a la excitación de electrones desde la banda de valencia a los niveles de energía correspondientes a los átomos de impureza. El átomo de impureza queda entonces ionizado, esto es, con carga negativa. La energía necesaria para este proceso, representada por E_a , corresponde a la diferencia de energía entre el borde superior de la banda de valencia y el nivel de energía del átomo ionizado (E_a). En este caso tendremos que $p > n$ y el semiconductor se denomina de tipo p .

Generalmente, el valor de la concentración de impurezas (donadoras o aceptoras) y su energía de excitación ($E_c - E_d$) ó ($E_a - E_v$), es tal que, a la temperatura ambiente, la concentración de portadores extrínsecos generados, n o p , es mucho mayor que la concentración de portadores intrínsecos, n_i . En el silicio por ejemplo, que tiene alrededor de 10^{22} átomos por cm^3 , la concentración de portadores intrínsecos es del orden de 10^{10} por cm^3 . Así pues, si se pretende que la concentración de portadores extrínsecos sea 10^4 veces mayor, es suficiente añadir una concentración de átomos de impureza en la proporción de 1 por cada 10^8 átomos de Si. La energía de excitación debe ser muy pequeña (alrededor de unas centésimas de electrón-voltio) con objeto de que a la temperatura ambiente todas las impurezas estén ionizadas. Este es el caso de las impurezas normalmente añadidas al silicio, como el B, P, As, etc. Por supuesto, la pureza del semiconductor de partida debe ser tal que la proporción de impurezas no deseadas sea muy inferior a las del dopaje, es decir de unas partes por billón, con objeto de que las propiedades de conducción estén determinadas exclusivamente por el dopaje. Estos requeri-

mientos de pureza, así como de la cristalinidad del material, son verdaderamente importantes para determinar con precisión las propiedades eléctricas de los semiconductores. Es evidente además que mediante la adición de impurezas de forma controlada se puede obtener un valor prefijado de antemano de la concentración de portadores, y por tanto de la conductividad del material, en un rango amplio de temperaturas.

En la fig. 1.11 se ha representado de manera cualitativa la variación de la concentración total del número de portadores en función de la temperatura para el caso del silicio dopado

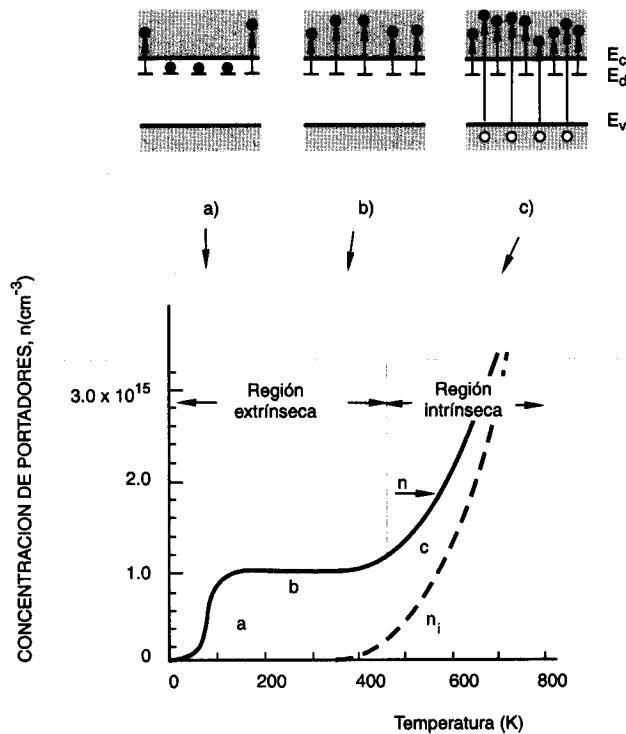


Fig.1.11. Variación con la temperatura de la concentración total de portadores para el Si tipo n, dopado con $N_d = 1 \times 10^{15} \text{ cm}^{-3}$ (línea continua). Se muestra también la variación de portadores, n_i , para el caso en el que semiconductor fuera intrínseco (línea a trazos). En la parte superior se indican los procesos de excitación correspondientes a cada rango de temperatura.

con átomos de arsénico, en una concentración de 10^{15} átomos por cm^3 (Si tipo n). A temperaturas muy bajas la proporción de átomos de As ionizados es muy pequeña por lo que la concentración de portadores (electrones) es también muy pequeña (tramo *a* de la curva). A temperaturas muy bajas, inferiores a los 100 K, comienza la ionización de los átomos de impureza, por lo que el valor de n aumenta rápidamente con la temperatura. Si se aumenta todavía más la temperatura llega un momento en que prácticamente todos los átomos de As quedan ionizados, con lo que n alcanza un valor estable. Esto ocurre desde una temperatura ligeramente superior a 100 K hasta unos 600 K, aproximadamente (tramo *b*). Por encima de esta temperatura comienza a dominar la ionización de los átomos de silicio aumentando de forma considerable el número de portadores intrínsecos, n_i . El valor de n aumenta de nuevo, iniciándose el comportamiento intrínseco (tramo *c*). Así pues, un semiconductor puede comportarse como intrínseco incluso cuando está dopado, para ello basta aumentar la temperatura suficientemente.

Como veremos en los próximos capítulos, el rango de operación útil de un semiconductor es el comprendido en el tramo *b* de la fig. 1.11, también llamado *rango de saturación*, ya que en esta región la concentración de portadores tiene un valor constante dentro de un margen amplio de temperaturas que incluye la temperatura ambiente. Además el valor de la concentración y el signo de los portadores es controlable a voluntad dentro de este rango. Basta para ello la adición de impurezas, aceptoras o donadoras, en la proporción debida. Es éste, quizás, uno de los aspectos más característicos de los semiconductores, ya que **mediante el proceso de dopaje es posible la prefijar la conductividad del material en un valor determinado**. Las técnicas utilizadas para la introducción de impurezas en el interior de un semiconductor son bastante complejas y serán descritas con detalle en el último capítulo de este tratado.

1.6. LEY DE ACCION DE MASAS

Según hemos visto anteriormente, los electrones en la banda de conducción y los huecos en la banda de valencia generados térmicamente adquieren una cierta energía cinética que les permite moverse en el interior del cristal. Sin embargo, el tiempo de vida de estos portadores no es infinito ya que una fracción importante de ellos está sometida constantemente a procesos de recombinación, mediante los cuales un electrón puede pasar a ocupar el nivel correspondiente a un hueco, desapareciendo tanto el electrón como el hueco y liberando al mismo tiempo una determinada energía. Estos procesos ocurren por ejemplo cuando un electrón pasa por las inmediaciones de un hueco, aunque existen también otros mecanismos posibles de recombinación. Si el semiconductor se halla en equilibrio térmico los procesos de recombinación están compensados a su vez por los de generación de nuevos pares electrón-hueco debidos a la excitación térmica. Supongamos un semiconductor intrínseco a una cierta temperatura, T . Si denominamos $R(T)$ y $G(T)$ al número de portadores que se recombinan y que se generan por unidad de tiempo, respectivamente, es evidente que en equi-

librio térmico tendremos (ver fig.1.12):

$$R(T) = G(T) \quad [1.5]$$

Las funciones $R(T)$ y $G(T)$ presentan una variación creciente con la temperatura, ya que a medida que aumenta T mayor es el número de electrones que se excita a la banda de conducción y lógicamente mayor es también el número de electrones que se desexcita a la banda de valencia. En cualquier caso, la ecuación [1.5] implica que en un semiconductor en equilibrio térmico la concentración de electrones en la banda de conducción y de huecos en la de valencia se mantiene constante con el tiempo.

Cuando un semiconductor se encuentra dopado, los procesos de generación y recombinación alteran profundamente la concentración de portadores (electrones y huecos) en el interior del semiconductor respecto al caso intrínseco. Así, cuando se añade una concentración N_d de impurezas donadoras en suficiente cantidad tendremos a la temperatura ambiente $n \approx N_d$, es decir, el valor de n se incrementa hasta igualar aproximadamente el valor de N_d . El aumento en la concentración de electrones en la banda de conducción origina a su vez una disminución en el número de huecos ya que existe una mayor probabilidad de procesos de recombinación. Algo parecido ocurre al añadir impurezas aceptoras a un semiconductor con una concentración N_a . En este caso el valor de p aumenta hasta el valor de N_a , es decir $p \approx N_a$, mientras que el de n disminuye. Consideremos el caso de un semiconductor tipo n con una concentración N_d de impurezas donadoras y queremos calcular la concentración de huecos en

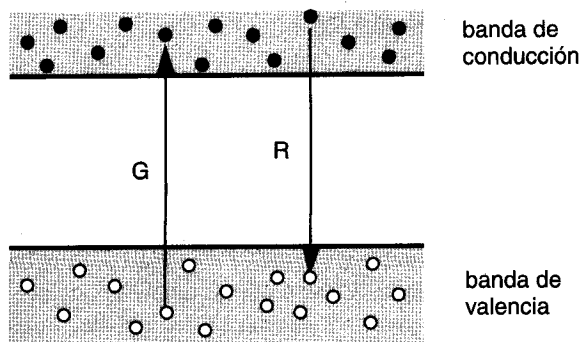


Fig. 1.12. Esquema de los procesos de generación, G , y de recombinación, R , de electrones y huecos en un semiconductor.

el semiconductor. Podemos suponer en este caso (aunque el razonamiento también es válido si se trata de un semiconductor tipo p) que la velocidad de recombinación, R , es proporcional al número de electrones presentes en la banda de conducción, n , y al de huecos en la banda de valencia, p , esto es:

$$R = r \cdot n \cdot p \quad [1.6]$$

siendo r una constante de proporcionalidad. Teniendo en cuenta la relación [1.5], y que para cualquier temperatura R y G han de ser constantes, la ecuación anterior conduce a:

$$n \cdot p = \text{cte} (T) \quad [1.7]$$

es decir, el producto np ha de ser una constante dependiente de la temperatura, exclusivamente, para una semiconductor determinado.

En realidad, la ecuación anterior es válida no sólo en el caso extrínseco sino también en el caso intrínseco. Es más, para un material semiconductor dado, bien sea puro o dopado, el valor de la constante en la ec. [1.7] ha de ser el mismo en un caso u otro. Para calcular el valor de esta constante podemos extender la ec. [1.7] al caso intrínseco en el que se cumple: $n = p = n_i$ (ec. 1.4). Se obtiene entonces: $n_i^2(T) = \text{cte}(T)$. Llevando el valor de la constante a la ec. [1.7] resulta la conocida *ley de acción de masas*:

$$n \cdot p = n_i^2(T) \quad [1.8]$$

La ecuación es completamente general, y por tanto válida tanto para los semiconductores intrínsecos como para los extrínsecos. A temperatura ambiente (300 K) el valor de n_i resulta ser $1.45 \times 10^{10} \text{ cm}^{-3}$ para el silicio y $1.79 \times 10^6 \text{ cm}^{-3}$ para el arseniuro de galio. En el caso del silicio puro, si se añade por ejemplo una concentración de impurezas donadoras, $N_d = 10^{14} \text{ cm}^{-3}$, supuesto que todas están ionizadas a la temperatura ambiente, tendremos para la concentración de electrones: $n = 10^{14} \text{ cm}^{-3}$, y para la de huecos: $p = n_i^2/n = (1.45 \times 10^{10})^2 / 10^{14} = 2.1 \times 10^6 \text{ cm}^{-3}$. Así pues, a medida que aumenta la concentración de electrones la de huecos se reduce en la misma proporción. Desde un punto de vista físico este proceso se explica si se tiene en cuenta que una pequeña fracción de los electrones procedentes de los átomos donadores contribuye a saturar una parte importante de los niveles o estados vacantes, disminuyendo por tanto la concentración de huecos.

Los datos anteriores muestran cómo es posible variar la concentración de electrones y huecos en un semiconductor mediante la adición de una proporción relativamente pequeña de átomos de impurezas. Efectivamente, en el caso que nos ocupa la concentración de impurezas añadidas (10^{14} átomos de impurezas por cm^3) representa una fracción despreciable frente a la

concentración de átomos de silicio del material (5×10^{22} átomos de Si por cm^3). Aún así, la concentración de electrones aumenta sobre la de huecos en varios órdenes de magnitud.

La ecuación [1.8] indica además que en un semiconductor extrínseco de tipo n, por ejemplo, con una concentración elevada de electrones en la banda de conducción, el número de huecos en la banda de valencia será muy pequeño. Se dice entonces que los electrones son los *portadores mayoritarios* y los huecos los *portadores minoritarios*. En semiconductores de tipo p, en condiciones análogas, tendremos la situación opuesta.

CUESTIONES Y PROBLEMAS

- 1.1 ¿Qué condiciones de pureza se exigirían a un material semiconductor como el Si para que pueda ser dopado con una concentración de 10^{14} átomos de impureza por cm^3 ?
- 1.2 Cuando se da el valor de la energía total de un electrón en la banda de valencia o de conducción, ¿cuál es el signo más apropiado para esta magnitud?, ¿por qué?. ¿Cuál sería la situación de un electrón con valor cero de energía?
- 1.3 Utilizando los datos del Sistema Periódico dar la configuración electrónica de los siguientes elementos: B, Si, Ge, As, Ga, Al, P y Sb. Asimismo, comparar los datos de valencia, radio iónico y potencial de ionización de estos elementos.
- 1.4 El arseniuro de galio se dopa con estaño. Si en la red del compuesto el estaño desplaza al galio, ¿cuál es el carácter de las impurezas?. ¿Qué tipo de semiconductor tendremos?
- 1.5 En el apartado 1.4 se explica que los huecos de la banda de valencia pueden participar en los procesos de conducción. ¿Quiere esto decir que dan lugar a una corriente adicional a la de electrones?. ¿Cuál sería el sentido de la corriente de huecos?
- 1.6 ¿Puede haber huecos en un metal?. Explicar bajo qué condiciones es conveniente introducir el concepto de hueco.

- 1.7** En un semiconductor dopado, ¿la concentración de portadores puede ser mayor que la concentración de impurezas añadidas?. Explicar las condiciones en que puede darse esta circunstancia.
- 1.8** Hacer un cálculo de la energía de ionización de las impurezas de silicio de tipo donador, utilizando un modelo de átomo hidrogenoide (núcleo positivo + un solo electrón). Comparar el resultado con la energía de la banda prohibida del Si.
- 1.9** ¿De qué factores cabe esperar que dependa la concentración de portadores intrínsecos en un material semiconductor?.
- 1.10** Si los procesos de generación y de recombinación de pares electrón-hueco se asimilan a una reacción química, ¿cuál serían los reactantes y los productos de la reacción?. Establecer la ley de masas en un semiconductor a partir de las leyes de equilibrio químico de la reacción.

CAPITULO II

PROCESOS DE TRANSPORTE DE CARGA EN SEMICONDUCTORES

En el capítulo anterior hemos visto los aspectos fundamentales de la estructura electrónica del enlace en los materiales semiconductores. Quizás lo más destacable de este estudio es la descripción del comportamiento del semiconductor en términos de la “teoría de bandas”, la cual permite explicar, al menos a nivel cualitativo, las características eléctricas de estos materiales. Sin embargo, los aspectos más cuantitativos de este problema, es decir, el número de cargas presentes en las bandas de valencia y de conducción y cómo varía su concentración con la temperatura u otros factores externos, aún no han sido abordados. Es más, interesa conocer también cómo contribuyen los portadores de carga a los procesos de conducción o lo que, de una manera más general, ha venido a denominarse *procesos de transporte*.

2.1. CALCULO DE LA CONCENTRACION DE PORTADORES A LA TEMPERATURA AMBIENTE

Antes de entrar en un cálculo detallado de la concentración de portadores y su variación con la temperatura es posible hacer un cálculo sencillo de su valor a la temperatura ambiente haciendo uso de la ley de acción de masas, ya estudiada en el capítulo anterior. Supondremos el caso más general de un semiconductor extrínseco dopado a la vez con una concentración N_d de impurezas donadoras y N_a de impurezas aceptoras. Según hemos visto

en el apartado 1.5, las impurezas donadoras cuando ceden un electrón quedan cargadas positivamente, mientras que las impurezas aceptoras cuando reciben un electrón quedan cargadas negativamente. A temperaturas muy bajas, inferiores a 100 K, solamente una fracción de esas impurezas se encuentran ionizadas. Si llamamos N_d^+ y N_a^- a la concentración de impurezas donadoras y aceptoras, respectivamente, que están ionizadas a una temperatura dada, la condición de *neutralidad de carga* exige que para esa temperatura se cumpla la relación:

$$N_d^+ + p = N_a^- + n \quad [2.1]$$

donde n y p es la concentración de electrones y huecos a esa temperatura. La ecuación anterior expresa que la concentración total de carga positiva ha de ser igual que la negativa, ya que en conjunto el semiconductor es neutro.

El cálculo de N_d^+ y N_a^- para cualquier temperatura es relativamente complejo. Sin embargo, cuando se trata de la temperatura ambiente normalmente todas las impurezas están ionizadas, por lo que tendremos entonces: $N_d^+ = N_d$ y $N_a^- = N_a$. La expresión anterior se reduce a:

$$N_d + p = N_a + n \quad (T = T_{amb}) \quad [2.2]$$

aunque en esta ecuación los valores de p y n no tienen por qué coincidir con los de la ecuación anterior. Supuesto n_i conocido, la ec. [2.2] junto con la [1.8] forma un sistema de dos ecuaciones con dos incógnitas que permite obtener separadamente los valores de n y p a la temperatura ambiente.

Para ilustrar mejor este cálculo supongamos un semiconductor extrínseco tipo n en el cual $N_a = 0$. Imponiendo la condición de que N_d sea suficientemente grande para que $n \gg p$, de las ecs. [1.8] y [2.2] resulta:

$$\begin{aligned} n &\approx N_d \\ p &\approx n_i^2/N_d \end{aligned} \quad (\text{tipo } n) \quad [2.3]$$

Análogamente para un semiconductor tipo p , con $N_d = 0$, la condición $p \gg n$ lleva a:

$$\begin{aligned} p &\approx N_a \\ n &\approx n_i^2/N_a \end{aligned} \quad (\text{tipo } p) \quad [2.4]$$

En semiconductores *compensados*, es decir, dopados a la vez con impurezas de tipo donador y aceptor, ocurre una situación intermedia. Así, si $N_a = N_d$, el semiconductor se com-

TABLA 2.1

CÁLCULO DE LA CONCENTRACIÓN DE PORTADORES EN DIFERENTES TIPOS DE SEMICONDUCTORES A LA TEMPERATURA AMBIENTE			
semiconductor		concentración de portadores	
tipo n		$n \approx N_d$	$p \approx n_i^2/N_d$
tipo p		$p \approx N_a$	$n \approx n_i^2/N_a$
compensado (*)	$N_d = N_a$	$n = p = n_i$	
	$N_d > N_a$ (*)	$n \approx N_d - N_a$	$p \approx n_i^2/(N_d - N_a)$
	$N_a > N_d$ (*)	$p \approx N_a - N_d$	$n \approx n_i^2/(N_a - N_d)$

(*) Con $|N_d - N_a| \gg n_i$

porta como intrínseco con $n = p$, mientras que si $N_d > N_a$, con la condición adicional de que el valor $N_d - N_a$ sea mucho mayor que n_i , entonces n viene dado por la diferencia $N_d - N_a$. Un resumen del cálculo de n y p a partir de la ley de acción de masas y del principio de neutralidad de carga para diferentes tipos de semiconductores viene dado en la tabla 2.1 (véase también el problema 2.1).

2.2. EFECTO DE LA TEMPERATURA EN LA CONCENTRACION DE PORTADORES

2.2.1. Distribución en energía de los portadores

Los electrones y huecos en un semiconductor tienen un comportamiento estadístico, de forma que su movimiento dentro de las bandas de conducción y de valencia se realiza con un cierto intercambio de energía entre ellos. Como resultado de este intercambio de energía, el conjunto de electrones en cada una de las bandas se distribuye entre los distintos niveles de la energía. Sin embargo, esta distribución no es uniforme en las diferentes regiones de energía dentro de la banda, ya que hay determinadas zonas o intervalos de energía donde la concentración de electrones puede ser mayor que en otras regiones de energía. La razón es debida a la diferencia en la cantidad de niveles disponibles para el electrón en cada intervalo de energía.

Incluso, en un intervalo determinado, la probabilidad de ocupación en un intervalo puede ser mayor que en otro.

Supongamos un intervalo pequeño de energía, ΔE , centrado alrededor del valor E de energía dentro de la banda de conducción. Si $\Delta n(E)$ representa la concentración total de electrones cuya energía está comprendida en ese intervalo, la densidad de electrones por unidad de intervalo de energía vendrá dado por el cociente $\Delta n(E)/\Delta E$. En el límite de intervalos infinitamente pequeños (pero conteniendo un gran número de niveles) este cociente tiende a un valor que denominaremos $n(E)$. Así pues, el producto $n(E)dE$ representa la fracción de electrones cuya energía está comprendida entre los valores E y $E+dE$. La función $n(E)$ se denomina también *función de distribución* de los electrones, ya que nos dice cómo se distribuyen los electrones en los diferentes intervalos infinitesimales de energía dentro de las bandas de energía del semiconductor.

Para calcular la densidad de electrones $n(E)$ en un semiconductor basta conocer la densidad de niveles disponibles en ese intervalo de energía y contar cuántos de ellos están ocupados por electrones. Así pues, podemos escribir $n(E)$ simplemente como:

$$n(E) = Z(E) \cdot f(E) \quad [2.5]$$

donde $Z(E)$ es la *densidad de estados posibles* en la banda (de conducción o de valencia), es decir, la concentración de estados o niveles por unidad de intervalo de energía que existe para un valor E de la energía y que son susceptibles de ser ocupados por los electrones, y $f(E)$ representa la *probabilidad de ocupación* de esos estados, o dicho en otras palabras, la fracción de estados o niveles que se encuentran ocupados.

En los semiconductores, al igual que en los metales, la probabilidad de ocupación de niveles por electrones está regida por la *estadística de Fermi-Dirac*. Esta estadística se aplica a partículas que cumplen el principio de exclusión de Pauli, como es el caso de los electrones. En la estadística de Fermi-Dirac, la función $f(E)$ está dada por:

$$f(E) = \frac{1}{1 + \exp [(E - E_F) / kT]} \quad [2.6]$$

siendo E_F un parámetro denominado *nivel de Fermi*, k la constante de Boltzmann y T la temperatura absoluta del sistema en consideración. La forma de la función $f(E)$, para distintas temperaturas incluida la del cero absoluto, está representada en la fig. 2.1 para un sistema de electrones. En este sistema el nivel de Fermi se ha situado arbitrariamente en 2.0 eV. Conviene señalar que el nivel de Fermi en un semiconductor es una magnitud que depende mucho de las características del semiconductor así como de otros factores externos, tales como la temperatura.

Obsérvese que, de acuerdo con [2.6] la función de probabilidad, $f(E)$, varía entre cero (estado vacante) y la unidad (estado ocupado). El nivel de Fermi, E_f , representa la energía para la cual la probabilidad de encontrar el electrón en ese nivel de energía vale $1/2$. Para niveles con una energía inferior a la del nivel de Fermi la probabilidad es siempre superior a $1/2$, mientras que para una energía superior la probabilidad es menor que $1/2$, decreciendo con la energía. La función de probabilidad es también dependiente de la temperatura. Así, a 0 K la probabilidad tiene un valor constante e igual a la unidad hasta la energía del nivel de Fermi. A partir de este valor la probabilidad disminuye abruptamente a cero. Esto quiere decir que a esta temperatura todos los niveles por debajo (o por encima) del nivel de Fermi se hallan ocu-

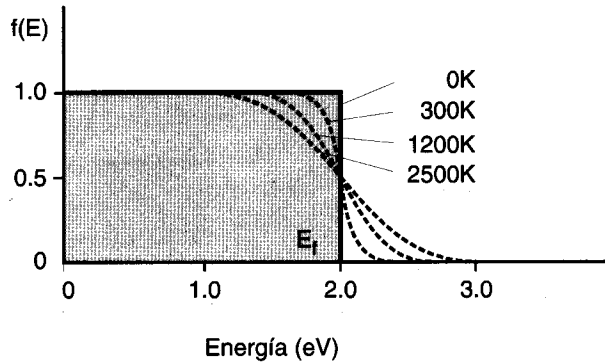


Fig.2.1. Función de probabilidad de Fermi-Dirac, $f(E)$, a diferentes temperaturas, para un sistema de electrones cuyo nivel de Fermi, E_f , está situado a 2.0 eV.

padados (o vacíos) de electrones. A temperaturas superiores a 0 K la función de distribución tiene una variación más suave, aunque la variación más pronunciada aparece alrededor del nivel de Fermi en un intervalo de energía del orden de kT .

En analogía con la ec. [2.5] podemos escribir para la función de distribución de los huecos en la banda de valencia, $p(E)$:

$$p(E) = Z(E) [1 - f(E)] \quad [2.7]$$

ya que un hueco representa un estado de energía vacante y por tanto la probabilidad de encontrar un hueco a la energía E vendrá dada por la función $[1 - f(E)]$.

El primer factor en las ecuaciones [2.5] y [2.7] representa, según se ha dicho, la densidad de estados o de niveles posibles para el electrón. Esta función depende también de la energía. En el caso de los electrones de un metal dentro de la banda de conducción, $Z(E)$ es una función creciente proporcional a $(E - E_c)^{1/2}$ tal como se representa en la fig. 2.2.a. Esto quiere decir que en el fondo de la banda de conducción no hay niveles susceptibles de ser ocupados por los electrones, mientras que a medida que subimos dentro de la banda la densidad de niveles aumenta de forma cuadrática con la energía. En los semiconductores intrínsecos y en los aislantes, $Z(E)$ puede ser una función bastante compleja y difícil de calcular. Generalmente se admite, por extrapolación del caso de los metales, que en las zonas próximas

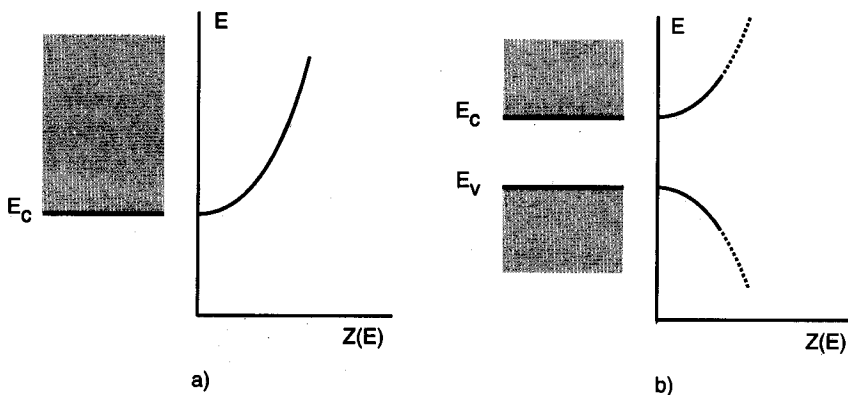


Fig. 2.2. Densidad de estados posibles, $Z(E)$, para los electrones en la banda de conducción de un metal (a) y en las proximidades de las bandas de valencia y conducción de un semiconductor intrínseco (b).

al tope de la banda de valencia, E_v , y al fondo de la banda de conducción, E_c , $Z(E)$ varía según las ecuaciones:

$$Z(E) = \frac{4\pi}{h^3} (2m_e^*)^{3/2} (E - E_c)^{1/2} \quad [2.8]$$

$$Z(E) = \frac{4\pi}{h^3} (2m_h^*)^{3/2} (E_v - E)^{1/2} \quad [2.9]$$

para las bandas de conducción y valencia, respectivamente (en las expresiones anteriores h es la constante de Planck y m_e^* y m_h^* son las masas efectivas de los electrones y huecos, respectivamente, definidas en la sec. 1.4.2). En la región intermedia de energía, zona de energía

prohibida, si el semiconductor es intrínseco la función $Z(E)$ es cero en toda la región (fig. 2.3). En cambio si el semiconductor es extrínseco, la función $Z(E)$ puede adoptar valores diferentes de cero en las proximidades de los niveles aceptores o donadores.

La fig. 2.3 muestra de forma gráfica el cálculo de la función de distribución para los electrones en el caso de un metal (fig. 2.3a), y para los electrones y huecos en el caso de un semiconductor (fig. 2.3b), utilizando las ecs. [2.5] y [2.7]. Como veremos más abajo, en este último caso el nivel de Fermi se halla situado generalmente entre las bandas de valencia y de conducción, mientras que en el caso de un metal el nivel de Fermi está situado dentro de la banda de conducción. Esta diferencia en la posición del nivel de Fermi es debida en cierto modo a que la banda de conducción del metal está mucho más poblada de electrones que la de

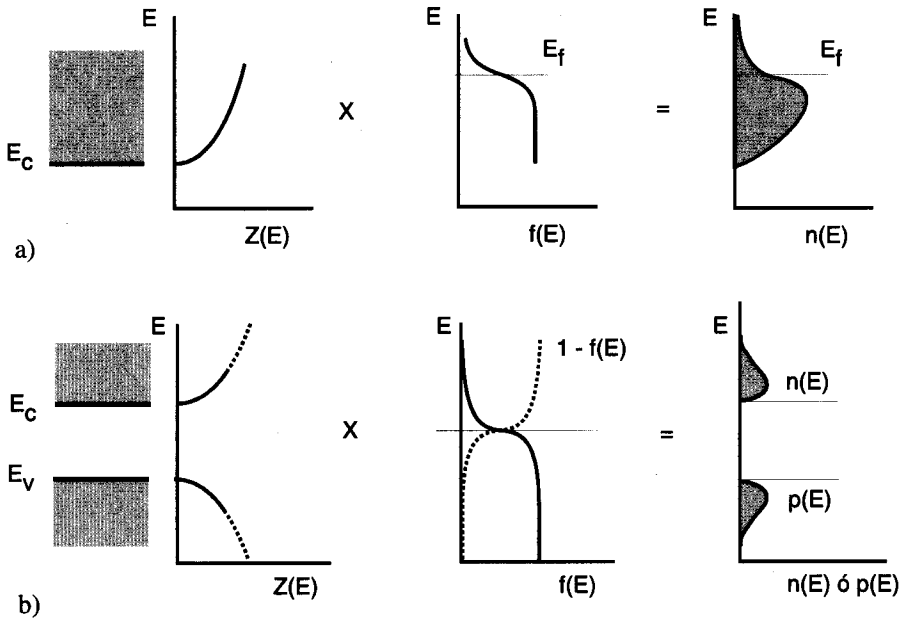


Fig. 2.3. Esquema del cálculo de la función de distribución de electrones, $n(E)$, en un metal (a) y de los electrones y huecos en un semiconductor (b).

un semiconductor. Nótese en la fig. 2.3a que en los metales la máxima concentración de electrones se encuentra en las proximidades del nivel de Fermi.

Conocida la función $Z(E)$ es posible calcular la concentración de electrones en la banda de conducción y de huecos en la banda de valencia mediante integración de las ecuaciones

[2.5] y [2.7]:

$$n = \int_{E_c}^{E_{\max}} Z(E) f(E) dE \quad [2.10]$$

$$p = \int_{E_{\min}}^{E_v} Z(E) [1 - f(E)] dE \quad [2.11]$$

siendo E_{\max} la energía del borde superior de la banda de conducción, y E_{\min} la energía del borde inferior de la banda de valencia. En el supuesto de ser conocida la energía del nivel de Fermi

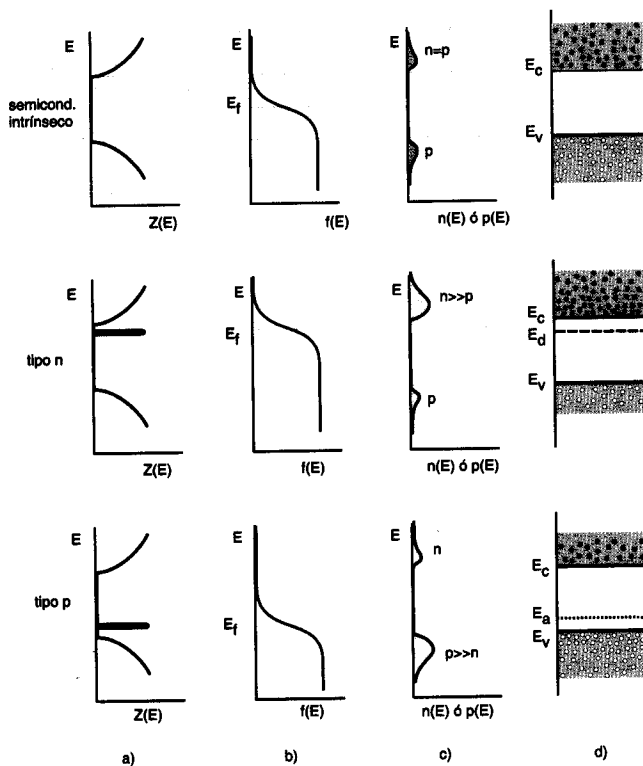


Fig.2.4. Curvas representativas de: a) la densidad de estados posibles, b) la probabilidad de ocupación y c) la distribución en energía de los portadores en diferentes tipos de semiconductores. En d) se da un esquema de la concentración de portadores. Nótese en b) la posición del nivel de Fermi en cada caso.

se puede calcular la concentración total de los electrones o de los huecos a partir de las expresiones anteriores. En realidad el cálculo sólo puede llevarse a cabo bajo ciertas aproximaciones, ya que $Z(E)$ sólo se conoce en las proximidades del fondo de la banda de conducción y en el borde superior de la banda de valencia.

En la fig. 2.4 se da un esquema de las funciones $Z(E)$ y $f(E)$, junto con la variación posible para $n(E)$ y $p(E)$, en diferentes tipos de semiconductores. En la columna de la izquierda, fig. 2.4a, se muestran las curvas correspondientes a la densidad de estados posibles, función $Z(E)$. Obsérvese que esta función es prácticamente la misma en todos los casos, excepto en la región de energías dentro de la banda prohibida, donde la función se hace cero si el semiconductor es intrínseco, o bien adopta valores finitos en ciertos niveles para tener en cuenta la existencia de estados donadores y aceptores (según sea el semiconductor n ó p) capaces de albergar electrones. La fig. 2.4b representa, asimismo, la función de probabilidad $f(E)$ que es también similar en todos los casos excepto en lo que respecta al nivel de Fermi, cuya posición varía según el tipo de semiconductor. En el siguiente apartado veremos cómo se determina en semiconductores de diferentes tipos la posición de este nivel. En semiconductores intrínsecos, E_f se encuentra aproximadamente en el centro de la banda prohibida. En los extrínsecos tipo n ó tipo p, este nivel se halla situado en las proximidades del fondo de la banda de conducción o del tope de la de valencia, respectivamente, aunque su posición exacta depende también de la temperatura. Finalmente, en la fig. 2.4c se ha representado las funciones $n(E)$ y $p(E)$. Según se aprecia las funciones de distribución $n(E)$ y $p(E)$ toman valores finitos,

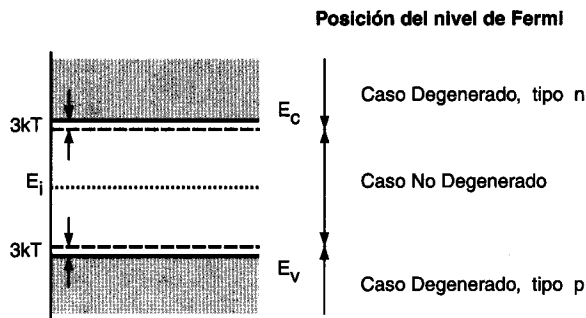


Fig.2.5. Esquema de los diferentes casos que se pueden presentar en un semiconductor según sea la posición del nivel de Fermi en relación con las bandas de energía. Se indica también la posición aproximada del nivel de Fermi intrínseco, E_i .

hasta un cierto máximo, por encima del borde de la banda de conducción y por debajo del tope de la banda de valencia (se supone en todos los casos que el semiconductor está a la temperatura ambiente), disminuyendo ambas funciones después del máximo a medida que nos alejamos de estos límites. Para un semiconductor intrínseco, $n(E)$ y $p(E)$ tienen la misma forma, con valores relativamente pequeños ya que el número de portadores generados térmicamente es muy bajo. En cambio las funciones $n(E)$ y $p(E)$ toman valores mucho más elevados cuando se trata de semiconductores extrínsecos (nótese la diferencia de concentraciones en la banda de valencia y de conducción en cada caso). Una representación cualitativa de cómo se distribuyen los portadores para cada tipo de semiconductor se da en la fig. 2.4d.

2.2.2. Cálculo de la concentración de portadores

Mediante la integración de las expresiones [2.10] y [2.11] es posible obtener una expresión general para la concentración total de portadores a cualquier temperatura (área de las curvas $n(E)$ y $p(E)$ en la fig. 2.4c). El cálculo es relativamente complejo, aunque se puede llegar a una ecuación simple bajo ciertas aproximaciones. Así, por ejemplo, para energías superiores o inferiores al nivel de Fermi en $3kT$, la función de Fermi-Dirac se puede aproximar a una función exponencial. En estas condiciones es posible hacer un cálculo aproximado de las integrales [2.10] y [2.11] para obtener n y p , aunque obviamente el cálculo está limitado al caso en que el nivel de Fermi del semiconductor se halle en el interior del “gap” y alejado de los bordes de la banda de valencia y de conducción en una distancia en energía superior a $3kT$ ¹. Esto ocurre con frecuencia en numerosos dispositivos electrónicos, y se dice entonces que el semiconductor es *no degenerado*. La condición de no-degeneración implica necesariamente que la concentración de electrones y huecos en las bandas de conducción y valencia no sea muy elevada y se presenta siempre que el dopaje del semiconductor no sea excesivamente alto y que la temperatura no sobrepase un cierto límite (fig. 2.5). Dentro de esta aproximación, las integrales [2.10] y [2.11] dan como resultado:

$$n = N_c \exp \left[- \frac{E_c - E_f}{kT} \right] \quad [2.12]$$

para la concentración total de electrones en la banda de conducción, y

$$p = N_v \exp \left[- \frac{E_f - E_v}{kT} \right] \quad [2.13]$$

para los huecos en la banda de valencia. Los parámetros N_c y N_v se denominan *densidad efectiva de estados* en la banda de conducción y de valencia, respectivamente. Su valor viene dado por las expresiones:

¹ **Nota:** El valor de kT a la temperatura ambiente es 0.0259 eV. Por tanto en un semiconductor no degenerado el nivel de Fermi debe estar como mínimo a una distancia de 0.078 eV de los bordes de la banda de valencia y de conducción, a la temperatura ambiente.

$$N_c = 2 (2\pi m_e^* kT/h^2)^{3/2} \quad [2.14]$$

$$N_v = 2 (2\pi m_h^* kT/h^2)^{3/2} \quad [2.15]$$

Tanto N_c como N_v se pueden interpretar como el límite máximo de estados posibles en las bandas de conducción y de valencia, respectivamente, de forma que si todos estuvieran ocupados su número representaría la concentración máxima posible de electrones y huecos, en cada caso. En realidad, si el semiconductor se considera no degenerado esta condición no se puede alcanzar, ya que implicaría que el nivel de Fermi estaría situado en los bordes de la banda de conducción o de valencia, según el caso. A la temperatura ambiente N_c y N_v tienen un valor muy próximo y es del orden de 10^{19} cm^{-3} (véase el problema 2.3). En cualquier caso, los valores de n y p se refieren siempre a la concentración total de portadores, tanto si se trata de semiconductores intrínsecos como extrínsecos.

La utilización de las ecs. [2.12] y [2.13] requiere fundamentalmente el conocimiento de la energía del nivel de Fermi. Su valor está determinado por el contenido de impurezas del semiconductor y otros parámetros del material. Es más, como veremos más adelante, las expresiones [2.12] y [2.13] se pueden utilizar también en sentido inverso, es decir, para determinar la posición del nivel de Fermi una vez que se conocen los valores de n ó p a través de la ley de acción de masas y el principio de neutralidad de carga (ecs. 1.8 y 2.1, respectivamente).

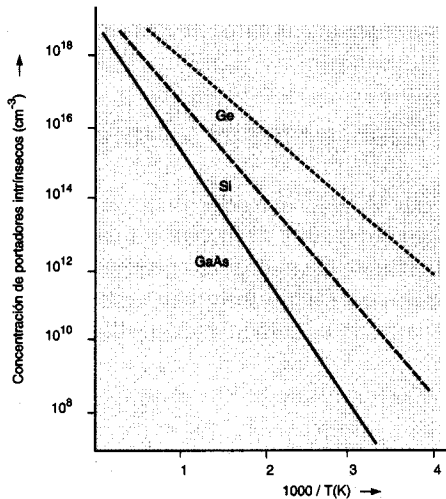


Fig.2.6. Dependencia de la concentración de portadores intrínsecos con el inverso de la temperatur. absoluta, para el germanio, silicio y arseniuro de galio.

Con objeto de obtener expresiones alternativas de n y p es conveniente introducir el *nivel de Fermi intrínseco*, E_i , que es el nivel de Fermi que tendría el semiconductor si fuera intrínseco, es decir, con una concentración de portadores igual a n_i . Más adelante veremos que el valor de E_i se halla situado aproximadamente a la mitad de la banda prohibida, según se esquematiza en la fig. 2.5. Supuesto conocido el valor de E_i , la concentración de portadores intrínsecos, n_i , se puede calcular a partir de las expresiones anteriores, haciendo en [2.12] $E_f = E_i$, obteniéndose:

$$n_i = N_c \exp \left[- \frac{E_c - E_i}{kT} \right] \quad [2.16]$$

Al mismo tiempo, la expresión [2.12] también se puede poner de la forma:

$$n = N_c \exp \left[- \frac{E_c - E_i}{kT} \right] \exp \left[\frac{E_f - E_i}{kT} \right]$$

con lo cual, utilizando la relación [2.16], resulta:

$$n = n_i \exp \left[\frac{E_f - E_i}{kT} \right] \quad [2.17]$$

El mismo procedimiento, usado para determinar la concentración de huecos lleva a:

$$p = n_i \exp \left[\frac{E_i - E_f}{kT} \right] \quad [2.18]$$

Las dos expresiones anteriores son de mucha utilidad, ya que el nivel de Fermi intrínseco se emplea muy a menudo como nivel de referencia en un esquema de bandas de energía. Esto ocurre sobre todo cuando se tienen semiconductores de la misma naturaleza pero con diferente dopaje. En estos casos, E_i es común a todos ellos, ya que su valor no depende del dopaje del semiconductor.

2.2.3 Concentración de portadores intrínsecos

Tratándose de semiconductores no degenerados, la concentración de portadores intrínsecos, n_i , se puede obtener fundamentalmente a partir de las ecs. [2.12] y [2.13], junto con la

ley de acción de masas (ec. 1.8):

$$n_i^2 = n p = N_c N_v \exp \left[- \frac{E_c - E_v}{kT} \right] \quad [2.19]$$

o bien, de forma más simple:

$$n_i = (N_c N_v)^{1/2} \exp \left[- \frac{E_g}{2kT} \right] \quad [2.20]$$

siendo $E_g = E_c - E_v$ la energía de la banda prohibida. La ec. [2.20] muestra que, independientemente del tipo de semiconductor, la concentración de portadores intrínsecos es una función exponencial decreciente de la anchura de la banda prohibida. La fig. 2.6 muestra la variación de la concentración de portadores en un semiconductor intrínseco con el inverso de la temperatura en una escala semilogarítmica, para tres semiconductores típicos: silicio, germanio y arseniuro de galio. Nótese que, de acuerdo con la expresión [2.20], en este tipo de representación la variación de n se convierte en una línea recta de pendiente negativa proporcional a E_g . Para el Si a temperatura ambiente resulta $n_i \approx 1.45 \times 10^{10} \text{ cm}^{-3}$. Como ya vimos en el capítulo anterior, esta concentración de portadores es muy baja cuando se la compara con la de los metales (del orden de 10^{22} ó 10^{23} cm^{-3}), por lo que el silicio puro a temperatura ambiente puede ser considerado como un aislante. Algo similar ocurre con el arseniuro de galio, e incluso también con el germanio.

2.3. DETERMINACION DEL NIVEL DE FERMI EN UN SEMICONDUCTOR

Según hemos visto anteriormente la determinación de la concentración de portadores en un semiconductor, ecs. [2.12] y [2.13] o bien [2.17] y [2.18], requiere el conocimiento de la posición de E_f . El valor de E_f puede obtenerse en cada caso concreto imponiendo la condición de neutralidad de carga, dada por la ec. [2.1], aunque sólo es posible obtener E_f de forma explícita dentro de ciertas aproximaciones.

Consideremos el caso simple de un semiconductor tipo n no compensado, es decir $N_a = 0$. Desde un punto de vista práctico es conveniente distinguir tres rangos de temperaturas:

- I) A **temperaturas** suficientemente **bajas** la concentración de portadores está controlada por el proceso de ionización de las impurezas, de forma que sólo una cierta fracción N_d^+ de los niveles donadores se encuentra ionizado. En este *rango*, denominado *extrínseco*, el cálculo de N_d^+ es complejo y por tanto es difícil conocer la posición del nivel

de Fermi. Sin embargo, se puede demostrar que en este caso el nivel de Fermi se encuentra situado aproximadamente a media distancia entre el fondo de la banda de conducción y la energía de los niveles donadores, aunque el valor exacto depende entre otros factores de la concentración de impurezas, la temperatura, etc. En cualquier caso, la concentración de electrones en la banda de conducción se puede aproximar por la expresión:

$$n \approx \left[\frac{N_d N_c}{2} \right]^{1/2} \exp \left[- \frac{E_d}{kT} \right] \quad [2.21]$$

siendo E_d la energía de los niveles donadores medida desde el fondo de la banda de conducción. De acuerdo con esta expresión a medida que aumenta la temperatura crece la concentración de portadores de una manera casi exponencial como consecuencia de la excitación de electrones desde los niveles donadores a la banda de conducción.

Una vez calculado n a través de la ecuación anterior, la concentración de huecos minoritarios en la banda de valencia se obtiene mediante la ec. [1.8]. Una ecuación similar a la ec. [2.21] se puede desarrollar también para el caso análogo de un semiconductor tipo p. La ec. [2.21] es comparable a la ec. [2.20], válida para semiconductores intrínsecos. De hecho, los procesos de excitación desde los niveles de impureza son completamente análogos a los de excitación desde la banda de valencia a la de conducción, la única diferencia es que la energía necesaria para la excitación es mucho menor en el primer caso, ya que $E_d \ll E_g$. Debido a ello, es evidente que a temperaturas bajas la concentración de portadores en un semiconductor está determinada por la ec. [2.21]. En esta región de temperaturas una representación del logaritmo de n en función del inverso de T debe dar una recta de pendiente negativa y proporcional a E_d .

- II) A **temperaturas intermedias** existe el llamado *rango de saturación* en el cual todas las impurezas se encuentran ionizadas siendo n esencialmente constante con la temperatura e igual a N_d (ec. 2.3). Esta característica hace que este rango de temperaturas, que puede extenderse entre 100 y 500 K, sea de gran utilidad cuando se desea controlar las propiedades de conducción de un semiconductor. La condición de $n = N_d$ permite además calcular de forma sencilla la posición del nivel de Fermi a partir de la ec. [2.12]. Se obtiene entonces:

$$E_c - E_f = kT \ln \left[\frac{N_c}{N_d} \right] \quad [2.22]$$

Igualmente, para un semiconductor tipo p, con $p = N_a$, se obtiene a partir de [2.13]:

$$E_f - E_v = kT \ln \left[\frac{N_v}{N_a} \right] \quad [2.23]$$

De estas expresiones se deduce que cuanto mayor sea la concentración de impurezas donadoras (o aceptoras) más próximo se halla el nivel de Fermi a la banda de conducción (o de valencia).

III) Finalmente, a medida que se eleva la temperatura por encima del rango de saturación, llega un momento en que la concentración de portadores intrínsecos se hace importante, de forma que su concentración puede alcanzar un valor sensiblemente superior al de los portadores puramente extrínsecos. En este rango de **temperaturas altas**, denomi-

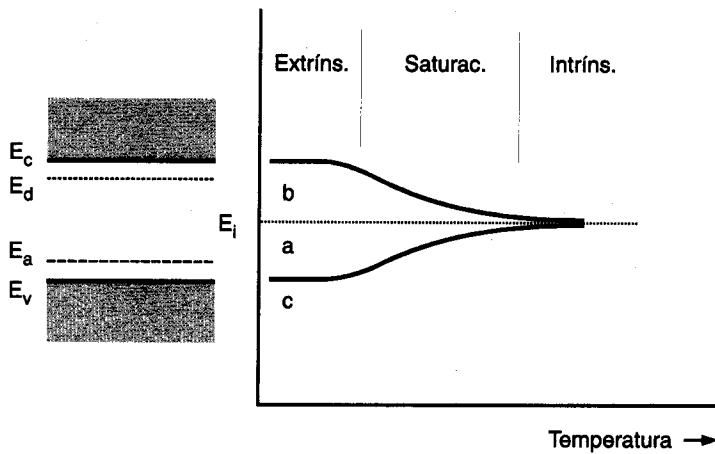


Fig.2.7. Posición del nivel de Fermi en función de la temperatura para diferentes tipos de semiconductores: a) intrínseco, b) tipo n, c) tipo p.

nado *rango intrínseco*, la concentración de portadores, determinada por la ec. [2.20], aumenta muy rápidamente con la temperatura. Del mismo modo, la posición del nivel de Fermi corresponde a la de un semiconductor intrínseco, E_i , y por tanto puede calcularse imponiendo en las ecs. [2.12] y [2.13] la condición de que en un semiconductor intrínseco el número de electrones en la banda de conducción ha de ser igual al de huecos en la banda de valencia, esto es $n = p$. Tendremos entonces a partir de estas ecuaciones:

$$E_f = E_i = \frac{E_c + E_v}{2} + \frac{kT}{2} \ln \frac{N_v}{N_c} \quad [2.24]$$

Dado que N_c y N_v son del mismo orden de magnitud, la expresión anterior pone de manifiesto que el nivel de Fermi intrínseco se halla prácticamente en el medio de la banda de energía prohibida mostrando una dependencia muy débil con la temperatura.

En la fig. 2.7 se presenta la variación del nivel de Fermi en los tres rangos de temperaturas indicados para distintos tipos de semiconductores. Como se puede apreciar, en el rango extrínseco el nivel de Fermi está situado a medio camino bien sea entre los niveles donadores y el fondo de la banda de conducción cuando el semiconductor es tipo n (curva b) o bien entre

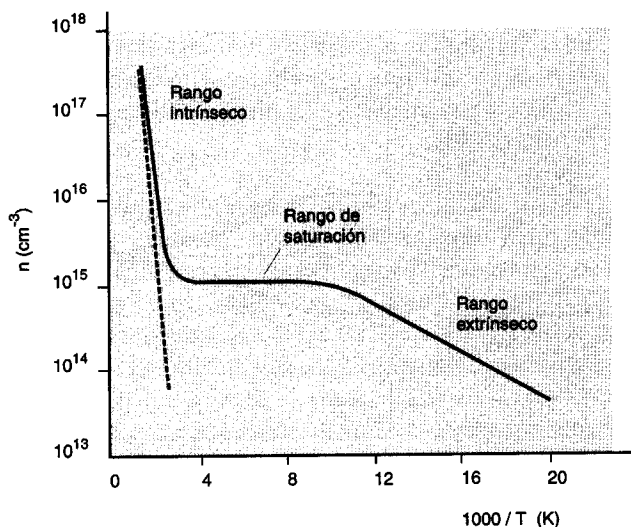


Fig.2.8. Esquema de la variación de la concentración de portadores en función del inverso de la temperatura absoluta en un semiconductor, dopado con una concentración de impurezas donadoras de 10^{15} cm^{-3} .

los niveles aceptores y el tope de la banda de valencia si es tipo p (curva c). A medida que aumenta la temperatura la variación del nivel de Fermi se rige por las ecuaciones [2.22] y [2.23], rango de saturación, las cuales predicen un descenso (o aumento) del nivel de Fermi si el semiconductor es de tipo n (o de tipo p). Finalmente cuando se alcanza el rango intrínseco a altas temperaturas el nivel de Fermi se aproxima al nivel intrínseco, E_i , es decir el nivel de Fermi que tendría el semiconductor si fuera intrínseco (curva a).

Como resumen de las ecuaciones desarrolladas en este apartado sobre el cálculo de la concentración de portadores en los materiales semiconductores, en la fig. 2.8 se muestra un

esquema de la variación de la concentración de electrones en función del inverso de la temperatura, dentro de los tres rangos estudiados, para un semiconductor tipo n dopado con una concentración de impurezas de 10^{15} cm^{-3} . En el rango extrínseco, la ec. [2.21] predice que en una escala logarítmica la variación debe ajustarse a una recta de pendiente negativa y proporcional a E_d . Cuando se alcanza el rango de saturación (zona intermedia de temperaturas) el valor de n se estabiliza en un valor igual a N_d . Según vimos en el capítulo anterior, este rango es el más apropiado para el funcionamiento de los dispositivos electrónicos ya que dentro de él la concentración de portadores es constante y además su valor se puede prefijar de antemano durante el proceso de preparación del material. Finalmente en el rango intrínseco la curva se aproxima de nuevo a una recta siguiendo las predicciones de la ec. [2.20]. La pendiente de esta recta debe ser proporcional a E_g , según ha sido ya mencionado.

2.4. PROCESOS DE CONDUCCION EN SEMICONDUCTORES

El transporte de carga en un semiconductor presenta más variedades que en un metal. Esto se debe a dos razones básicas:

- I) Los semiconductores generalmente poseen dos tipos de portadores, los cuales se mueven en sentidos opuestos cuando se aplica un campo eléctrico, y
- II) En un semiconductor puede haber fuertes variaciones locales en la concentración de portadores, las cuales a su vez dan lugar a desplazamientos locales de carga a través de un mecanismo de difusión. Este mecanismo de transporte, que tiene una gran importancia en el funcionamiento de los diodos y transistores, será descrito en el siguiente apartado.

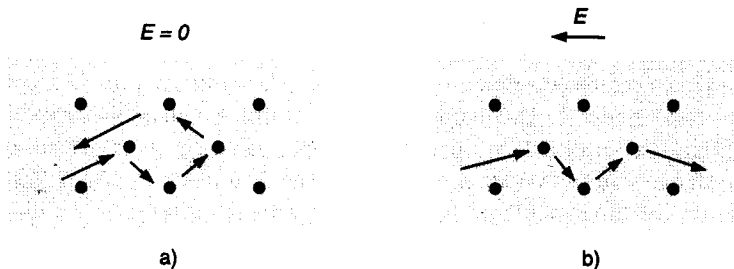


Fig. 2.9. Esquema del movimiento de un electrón en la banda de conducción a través de la red de átomos del semiconductor, a) en ausencia de campo eléctrico, b) con un campo eléctrico aplicado, E .

Consideremos un semiconductor en equilibrio térmico, a una temperatura T , sin aplicar un campo eléctrico. Los electrones y huecos poseen en estas condiciones una cierta energía cinética como consecuencia de su excitación térmica a los niveles de la banda de conducción o de valencia. Según se mencionó en el capítulo anterior (apartado 1.4.2), la energía cinética de los electrones (o huecos) está determinada por la diferencia entre su energía total y la del borde de la banda de conducción (o valencia, cambiada de signo). Debido a esta energía cinética, los electrones y huecos se hallan en continuo movimiento a través de la red, interaccionando con los átomos. En los procesos de interacción con los átomos, los portadores pueden perder o incluso ganar cierta cantidad de energía, aunque en conjunto la función de distribución de los electrones y huecos, y por tanto su energía cinética media se mantiene constante. Para hacernos una idea del orden de magnitud de la velocidad media de los electrones, $\langle v \rangle$, podemos hacer un cálculo sencillo suponiendo que toda la energía térmica transferida por la red, $1/2 kT$ por cada grado de libertad, se convierte en energía cinética de traslación. Para un electrón moviéndose con tres grados de libertad tendremos por tanto:

$$\frac{1}{2} m_e \langle v^2 \rangle = \frac{3}{2} kT$$

donde m_e es la masa efectiva para la conductividad, con un valor diferente a la masa efectiva que aparece en el cálculo de la densidad efectiva de estados (m_e^*). A la temperatura ambiente (≈ 300 K) la ecuación anterior arroja un valor de $\langle v \rangle \approx 10^5$ ms⁻¹ para los electrones moviéndose en la banda de conducción del silicio o del arseniuro de galio. Aunque la velocidad media puede parecer elevada, hay que tener en cuenta que los electrones cambian constantemente su dirección debido a las colisiones con los átomos de la red y con las impurezas del dopaje. Resulta así un movimiento completamente aleatorio que no supone un desplazamiento neto de carga (fig. 2.9.a). La distancia media entre colisiones se caracteriza por el *recorrido libre medio*, y el tiempo entre colisiones se denomina *tiempo libre medio*, t_c . Este último parámetro resulta del orden de 10^{-12} s.

Cuando se aplica un campo eléctrico E en una dirección determinada tenemos una situación diferente ya que el electrón está sujeto a una aceleración debida a una fuerza de magnitud qE en dirección opuesta al campo. Sin embargo, una parte de la energía ganada del campo eléctrico es cedida a la red cristalina debido a las colisiones de los electrones con los átomos que están en reposo. Como consecuencia de ello, los electrones experimentan en su conjunto un arrastre en dirección opuesta al campo, ya que a la velocidad $\langle v \rangle$, debida al movimiento térmico, se superpone ahora la velocidad de arrastre, v_e (fig. 2.9b). Podemos de nuevo hacer un cálculo sencillo de la velocidad media de arrastre de los electrones, $\langle v_e \rangle$, igualando el impulso mecánico del campo eléctrico entre dos colisiones, dado por $-q E t_c$, a la variación media de la cantidad de movimiento, $m_e \langle v_e \rangle$. Resulta así:

$$\langle v_e \rangle = - \frac{q t_c}{m_e} E \quad [2.25]$$

El factor de proporcionalidad entre $\langle v_e \rangle$ y E es una nueva magnitud, característica del movimiento considerado para los electrones, denominada *movilidad*, μ_e . Para el mecanismo descrito de colisión de los electrones con los átomos de la red, esta magnitud está dada por:

$$\mu_e = \frac{q t_c}{m_e} \quad [2.26]$$

De acuerdo con esta expresión, la ecuación [2.25] resulta:

$$\langle v_e \rangle = - \mu_e E \quad [2.27]$$

con una ecuación análoga (con signo positivo) para los huecos. Uno de los factores que más afecta el valor de la movilidad es la temperatura del semiconductor, ya que la temperatura produce un aumento de la amplitud de vibración de los átomos de la red, lo cual a su vez hace que el tiempo libre medio entre colisiones sea cada vez más pequeño. Esto da lugar a una disminución de la movilidad con la temperatura. Sin embargo, en semiconductores muy dopados la interacción de los electrones con las impurezas puede dominar sobre otros mecanismos de colisión produciendo una variación de t_c con la temperatura más compleja.

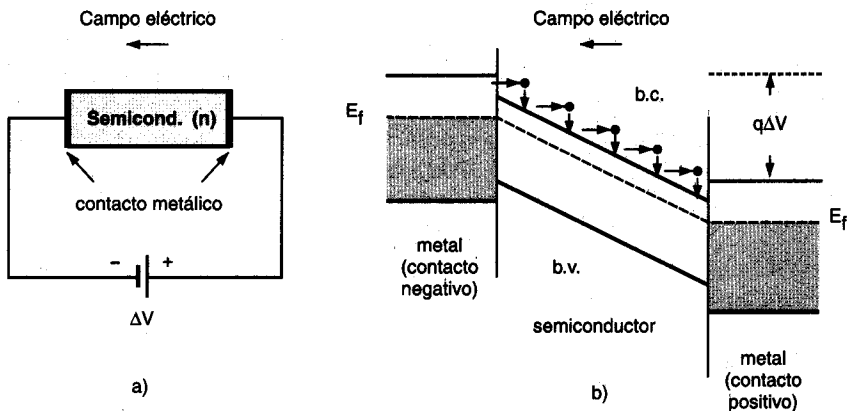


Fig. 2.10. Esquema del movimiento de portadores en un semiconductor: a) La aplicación de una diferencia de potencial en sus extremos, ΔV , da lugar a un desnivel de las bandas de energía en la cantidad $q\Delta V$. b) Movimiento de los electrones en la banda de conducción por la acción de la caída de potencial en el semiconductor.

Al estudiar el fenómeno de conducción, la presencia de un campo eléctrico constante en el interior del semiconductor normalmente es debida a la aplicación de una diferencia de potencial, ΔV , en sus extremos a través de sendos contactos metálicos (fig. 2.10a). Esto implica que los electrones en el lado positivo tienen una energía potencial más baja, de valor igual a $-q\Delta V$, que en el lado negativo. Según vimos en el apartado 1.4.2, en un diagrama de energías esta situación se representa por una inclinación uniforme de las bandas de energía del semiconductor en esa misma cantidad, tal como se indica en la fig. 2.10b. Por contra, las bandas de energía de los contactos metálicos quedan inalteradas, ya que se supone que el campo eléctrico en el interior del metal es nulo y por tanto no existe caída de potencial. Dejando aparte el transporte de carga desde el metal al semiconductor (este problema será tratado con detalle en el cap. IV), se puede visualizar en un esquema de energías el movimiento de arrastre de los electrones entre dos colisiones dentro del semiconductor mediante un desplazamiento horizontal dentro de la banda de conducción, implicando con ello una ganancia en energía cinética a costa de una pérdida en energía potencial. En el proceso de colisión hay una pérdida de energía cinética (que se transfiere a la red en forma de calor dando lugar al efecto Joule), cayendo el electrón (en energía) prácticamente al fondo de la banda de conducción (flecha vertical). Después de esta pérdida de energía el electrón se acelera de nuevo ganando energía del campo hasta que sufre una nueva colisión. Para los huecos ocurre una situación similar aunque viajando en sentido opuesto y aproximándose tras cada colisión al tope de la banda de valencia.

Es importante mencionar que la velocidad de arrastre de los electrones puede ser muy diferente a la de los huecos. Según se ha visto en apartados anteriores, el desplazamiento de los huecos implica cambios en los enlaces interatómicos mientras que los electrones se mueven con relativa libertad dentro de la nube electrónica a través de los átomos. Por ello es de esperar, como ocurre en el caso del silicio, que la movilidad de los huecos sea mucho menor que la de los electrones. Los valores experimentales de la movilidad obtenidos para los semiconductores típicos, Si, Ge y GaAs vienen recogidos, junto con otros datos característicos, en la tabla 2.2.

Para un conjunto de electrones desplazándose en el interior de un semiconductor con una velocidad media de arrastre determinada, $\langle v_e \rangle$, la densidad de corriente J_e , es decir la corriente eléctrica I_e por unidad de área viene dada (en módulo) por:

$$J_e = -q n \langle v_e \rangle = q n \mu_e E \quad [2.28]$$

Análogamente para los huecos, si su concentración es p y la movilidad μ_h , tendremos:

$$J_h = q p \mu_h E \quad [2.29]$$

La densidad de corriente total, $J = J_e + J_h$, debida a ambas contribuciones será por tanto:

$$J = J_e + J_h = (q n \mu_e + q p \mu_h) E \quad [2.30]$$

o bien:

$$J = \sigma E \quad [2.31]$$

siendo σ la *conductividad* del semiconductor definida por el factor entre paréntesis de la ecuación [2.30]:

$$\sigma = q n \mu_e + q p \mu_h \quad [2.32]$$

TABLA 2.2

PARAMETROS CARACTERÍSTICOS DE ALGUNOS SEMICONDUCTORES			
	Si	Ge	GaAs
Átomos/cm ³	5.0x10 ²²	4.4x10 ²²	4.4x10 ²²
Densidad (gr/cm ³)	2.3	5.3	5.3
Permitividad relativa	11.9	16.0	13.1
Densidad efectiv. de estados (cm ⁻³)			
N_c	2.8x10 ¹⁹	1.0x10 ¹⁹	4.7x10 ¹⁷
N_v	1.0x10 ¹⁹	6.1x10 ¹⁹	7.0x10 ¹⁸
Energía banda prohibida (eV)	1.12	0.66	1.43
Conc. portadores intrín., T_{amb} (cm ⁻³)	1.5x10 ¹⁰	2.4x10 ¹³	1.8x10 ⁶
Masa efectiva (conduc.)			
m_e/m_o	0.26	0.12	0.09
m_h/m_o	0.38	0.23	---
Movilidad (cm ² V/s)			
μ_e	1.5x10 ³	3.9x10 ³	8.5x10 ³
μ_h	0.6x10 ³	1.9x10 ³	0.4x10 ³

La conductividad es también un parámetro que depende no sólo de las características del semiconductor, sino también de la temperatura. Según hemos visto, la concentración de portadores aumenta exponencialmente dentro de determinados rangos de temperatura. A su vez la movilidad puede disminuir con la temperatura, aunque no lo suficiente como para compensar el aumento debido a la concentración de portadores. Esto hace, en conjunto, que la conductividad de un semiconductor en el rango intrínseco aparezca como una función creciente con la temperatura, justo al contrario de lo que sucede con los metales en los que, debido a que en ellos la concentración de portadores es constante, la conductividad disminuye con la temperatura.

2.5 PROCESOS DE DIFUSION

Los procesos de difusión son muy frecuentes en los semiconductores y ocurren cuando, por cualquier circunstancia, se produce una variación de la distribución portadores de un punto a otro en el interior del material. Así, por ejemplo, supongamos el caso hipotético de un semiconductor extrínseco preparado con una distribución inhomogénea de impurezas. Inicialmente la concentración de portadores, para cualquier intervalo de energía, será mayor en aquellas zonas donde el contenido de impurezas sea también mayor, ya que a temperatura ambiente todas las impurezas estarán ionizadas. Otro ejemplo en el cual aparece también un exceso de portadores ocurre cuando el semiconductor se ilumina en un punto determinado con radiación de energía suficiente para excitar localmente electrones desde la banda de valencia a la de conducción.

En todos estos casos, la variación local de la concentración de portadores de un punto a otro origina un proceso de difusión, mediante el cual los portadores de carga se mueven desde las zonas donde la concentración es mayor a aquellas donde es menor, tendiendo a igualar la concentración en todos los puntos del cristal. El fenómeno de la difusión tiene su origen en el movimiento aleatorio de los electrones dentro del semiconductor durante el cual están continuamente intercambiando su energía. Quizás, el símil más parecido es el de la difusión de un fluido a través de una membrana que separa dos recintos de diferentes concentraciones, o el caso de la difusión del calor de un punto a otro debida a la diferencia de temperaturas. De hecho, las ecuaciones que rigen el fenómeno de la difusión son muy similares en todos estos casos. Así, por ejemplo, si la concentración de huecos en un semiconductor es una función de la posición x en el interior del semiconductor, es decir $p = p(x)$, existe entonces una corriente de huecos cuyo valor por unidad de superficie, J_h , es proporcional a la derivada de la función $p(x)$, es decir:

$$J_h = -q D_h \frac{dp}{dx} \quad [2.33]$$

Igualmente, si en el semiconductor existe una variación $n(x)$ de la concentración de electrones tendremos:

$$J_e = q D_e \frac{dn}{dx} \quad [2.34]$$

donde D_h y D_e son constantes características del material, y se conocen como los *coeficientes de difusión* para los huecos y electrones, respectivamente. Debido a que los portadores se mueven en la dirección de los puntos de mayor a menor concentración, la corriente de difusión en el caso de los huecos tiene signo opuesto al de la derivada (ec. 2.33). Cuando existe difusión de electrones y huecos en presencia de un campo eléctrico E aplicado en la dirección del eje x , la corriente será en cada caso:

$$J_e = q n \mu_e E + q D_e \frac{dn}{dx} \quad [2.35]$$

$$J_h = q n \mu_h E - q D_h \frac{dp}{dx} \quad [2.36]$$

Para un semiconductor no degenerado, se puede demostrar además que cuando los movimientos de los portadores por arrastre del campo eléctrico y por difusión procede con el mismo mecanismo de interacción con los átomos de la red entonces se cumple la *ecuación de Einstein* para las movilidades:

$$\frac{D_h}{\mu_h} = \frac{D_e}{\mu_e} = \frac{kT}{q} \quad [2.37]$$

Al cociente $V_T = kT/q$ se le denomina *voltaje equivalente de la temperatura*. Para la temperatura ambiente, $V_T = 0.0259$ Voltios. La ec. [2.37] muestra un resultado esperado, es decir, la relación de proporcionalidad entre el coeficiente de difusión y la movilidad de los portadores. Como veremos en el siguiente capítulo, las corrientes de difusión tienen una gran importancia en el funcionamiento de los diodos y en otros dispositivos semiconductores. Generalmente, los dispositivos electrónicos están formados por la unión de dos o más semiconductores. Se trata por tanto de sistemas inhomogéneos en los que la concentración de portadores puede variar de un punto a otro, dando lugar a corrientes de difusión.

2.5.1. Semiconductores con dopaje no uniforme: curvatura de las bandas de energía

Desde un punto de vista energético, el proceso de la difusión tiene también un gran interés, ya que puede dar lugar a un efecto de curvatura de las bandas de energía y a la aparición de campos eléctricos en el interior del semiconductor. Consideremos de nuevo el caso de

un semiconductor de tipo n con dopaje no homogéneo, en el que la concentración de impurezas donadoras varía a lo largo de la coordenada x , según se indica en el esquema de la figura 2.11a. Inicialmente, antes de alcanzarse el equilibrio, la concentración de electrones en la banda de conducción tendrá también una variación similar a lo largo del eje x . Esto quiere decir que en la región izquierda del semiconductor, el punto de coordenada x_1 por ejemplo, la función de distribución de los electrones, $n(E)$, toma valores más elevados que en la región derecha, en la coordenada x_2 . Asimismo, el nivel de Fermi variará su altura de acuerdo con la concentración de portadores en cada punto, con una posición más próxima a la banda de conducción en los puntos donde la concentración de portadores es mayor (fig. 2.11.b).

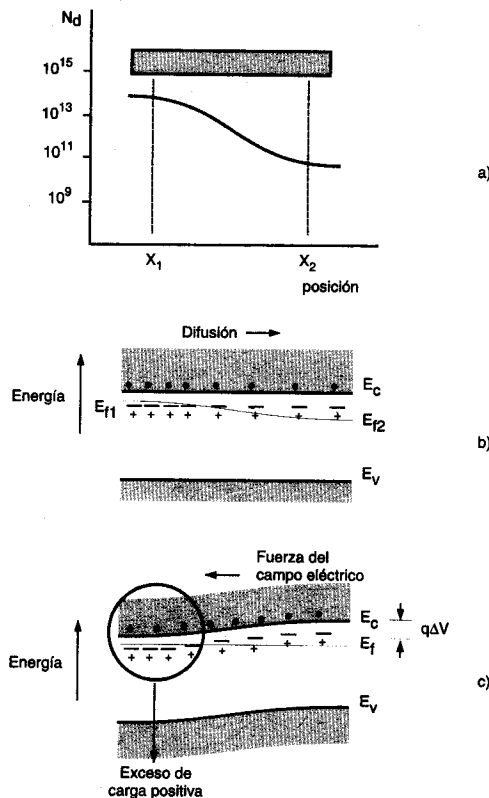


Fig. 2.11. Esquema del proceso de difusión en una barra semiconductor con dopaje inhomogéneo, tipo n: a) Variación de la concentración de impurezas a lo largo de la barra. b) Bandas de energía antes de establecerse el equilibrio (obsérvese que la posición del nivel de Fermi no es constante). c) Situación después de alcanzarse el equilibrio.

Debido a la diferente concentración en cada punto, antes de alcanzarse el equilibrio existirá una corriente de difusión de electrones cuyo valor vendrá dado por la ecuación [2.34]. Dado que no existe una fuente externa de carga, esta corriente de difusión no puede durar indefinidamente. En un gas de partículas sin carga, el proceso de difusión se mantendría hasta que se igualaran las concentraciones en todos los puntos. En un semiconductor, por el contrario, la difusión cesa mucho antes, ya que el movimiento de electrones hacia un lado de la barra semiconductor origina una acumulación de carga negativa en ese lado, dejando un exceso de carga positiva en el lado opuesto de la barra. Esta carga positiva, también denominada *carga espacial*, es carga fija, ya que es debida a la carga de los iones de las impurezas del semiconductor que quedan sin compensar (fig. 2.11c). Después de alcanzarse el equilibrio, la presencia de estas cargas origina a su vez un campo eléctrico E que tiende a oponerse al movimiento de electrones por efecto de la difusión. En realidad se trata de una situación de equilibrio dinámico, donde la corriente de difusión es compensada por otra corriente igual y de sentido contrario debida al campo eléctrico originado por la carga espacial.

Es fácil calcular el potencial eléctrico asociado al campo creado por la carga espacial. En el caso que nos ocupa, si $n = n(x)$ representa la concentración de electrones en cada punto x del semiconductor, la corriente total debido a ambos fenómenos de difusión y de arrastre por el campo eléctrico vendrá dado por la ec. [2.35]. La condición de equilibrio implica además $J_e = 0$, por tanto:

$$E = - \frac{D_e}{\mu_e} \frac{1}{n} \frac{dn}{dx} \quad [2.38]$$

Teniendo en cuenta que $dV = -Edx$, tendremos:

$$dV = \frac{D_e}{\mu_e} \frac{dn}{n}$$

Si en la coordenada x_1 la concentración de portadores es n_1 , y en x_2 la concentración es n_2 (fig. 2.11a), la ecuación anterior se puede integrar resultando:

$$\Delta V = V_2 - V_1 = \frac{D_e}{\mu_e} \ln \frac{n_2}{n_1} \quad [2.39]$$

La ecuación [2.39] indica que la diferencia de potencial entre dos puntos con diferente concentración de portadores sólo depende de la concentración relativa de portadores en esos dos puntos. Evidentemente, el signo del potencial es tal que tiende a oponerse a la corriente de difusión (en el ejemplo de la fig. 2.11 el potencial es tanto más positivo cuanto más a la izquierda del semiconductor).

Como consecuencia del campo eléctrico creado después del proceso de difusión, los portadores poseen una energía potencial adicional a la que ya tienen por el hecho de moverse

en la banda de conducción o de valencia. Así, en el ejemplo considerado en la fig. 2.11 un electrón situado en la zona izquierda del semiconductor (donde hay exceso de carga positiva) tendrá una energía potencial menor que en la zona de la derecha. Esta situación queda descrita de forma sencilla mediante una curvatura de las bandas de forma que la diferencia de energía potencial en sus extremos sea igual a $q\Delta V$, tal como se representa en la fig. 2.11c. Hay que recordar que, según se mencionó en el capítulo anterior, la existencia de un campo eléctrico en el interior del semiconductor trae consigo una curvatura de las bandas de energía, siendo la pendiente de la curva en cada punto proporcional al valor del campo eléctrico cambiado de signo (véase sec. 1.4.3). En realidad, la curvatura de las bandas es consecuente con el cambio de la energía de los electrones, de forma que la función de distribución, $n(E)$, en cada punto x del semiconductor queda también desplazada en una proporción igual a la de las bandas. Si se tratara de un semiconductor tipo p con una distribución similar, la curvatura de las bandas de energía sería entonces descendiente hacia la derecha.

2.5.2. Constancia del nivel de Fermi

Es importante mencionar que el proceso de difusión no implica cambios en la energía total del sistema siempre que el semiconductor se encuentre aislado y en equilibrio térmico. En consecuencia, la energía del nivel de Fermi, E_F , se debe mantener constante en el interior de todo el semiconductor, tal como se indica en la fig. 2.11c. Efectivamente, es fácil demostrar que la energía $q\Delta V$, asociada a la curvatura de las bandas después de alcanzarse el equilibrio es exactamente igual a la diferencia de energía de los niveles de Fermi, $E_{F2} - E_{F1}$, correspondientes a los puntos x_2 y x_1 del semiconductor (véase problema 2.10). Por tanto, el desplazamiento de las bandas de energía da lugar a un corrimiento paralelo del nivel hasta que queda horizontal. Nótese en la fig. 2.11c que la distancia del nivel de Fermi a la banda de conducción en cada punto sigue siendo la misma que antes de producirse la curvatura de las bandas, ya que la función de distribución, $n(E)$, queda prácticamente inalterada.

Este resultado es aplicable no sólo al caso de un semiconductor único sino a cualquier sistema formado por la unión de materiales de diferentes tipos, conductores o no, siempre que el sistema esté aislado y en equilibrio termodinámico. De hecho se puede establecer como un principio fundamental que **en condiciones de equilibrio el nivel de Fermi en todo el interior de un material aislado (o de un grupo de materiales en contacto), es decir, sin campo eléctrico aplicado, tiene un valor constante**. Como veremos mas adelante, este principio tiene una gran aplicación en el estudio del comportamiento de los dispositivos semiconductores.

El principio de igualdad del nivel de Fermi en todos los puntos del interior del semiconductor no debe sorprendernos. Evidentemente, si el sistema de electrones se encuentra en equilibrio térmico, niveles con la misma energía deben tener también la misma probabilidad de ocupación en cualquier punto del cristal. O a la inversa, niveles que tengan la misma probabilidad (como es el caso del nivel de Fermi, con probabilidad 1/2) han de tener la misma

energía, independientemente de la posición en el interior del semiconductor. Se puede, pues, concluir que **un sistema en el que la distribución de electrones se encuentra inicialmente fuera del equilibrio evoluciona hasta que el nivel de Fermi es constante en el interior del semiconductor.**

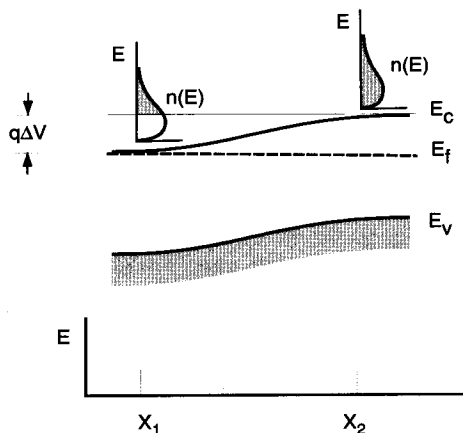


Fig. 2.12. Esquema de la función de distribución de los electrones en un semiconductor con dopaje inhomogéneo (según los datos de la fig. 2.11), después de alcanzarse el equilibrio.

De acuerdo con estos postulados, se puede decir que la curvatura en las bandas de energía tiende a mantener constante en cada punto del cristal la probabilidad de ocupación de cada nivel de energía permitido. Siguiendo con el ejemplo de la fig. 2.11, es fácil demostrar que en la región izquierda del semiconductor -coordenada x_1 - la concentración de electrones situados en la cola de la función de distribución, $n(E)$, con energía superior a $q\Delta V$ (zona rayada), debe ser ligeramente mayor (aunque bajo ciertas aproximaciones se puede considerar prácticamente igual) que la concentración de electrones en cualquier otro punto situado a la derecha del cristal -coordenada x_2 - (fig. 2.12). Son estos electrones, con energía superior a $q\Delta V$ por encima de la banda de conducción, los que pueden moverse libremente hacia la derecha en el proceso de difusión. En el equilibrio esta corriente de difusión se ve compensada por la corriente de electrones que se desplaza en sentido opuesto empujada por el campo eléctrico presente en el semiconductor.

2.6. PROCESOS DE INYECCION DE PORTADORES

Hasta ahora hemos considerado que el semiconductor se encuentra en una situación de equilibrio térmico, en el cual la tasa de recombinación está siempre compensada por la de

generación manteniéndose constantes las concentraciones n y p . Sin embargo, no siempre se tienen situaciones de equilibrio ya que existen muchos procesos de generación de portadores que tienden a aumentar los valores de n y p sobre los de equilibrio térmico. Estos procesos de generación, distintos al térmico, son, por ejemplo, los debidos a la excitación de portadores cuando el material se ilumina con radiación electromagnética de energía suficiente (fig. 2.13).

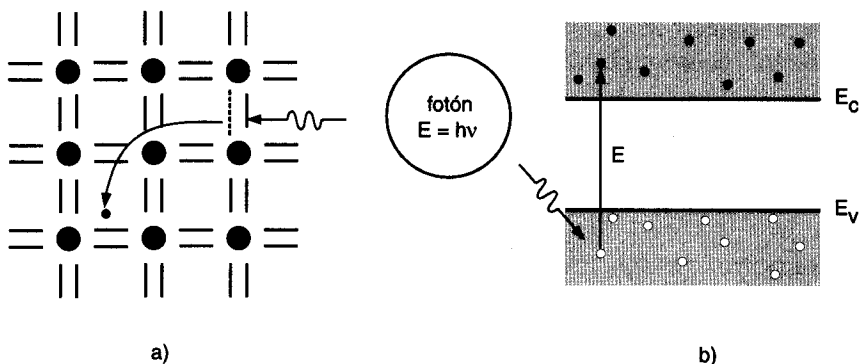


Fig.2.13. Proceso de inyección de portadores mediante excitación con radiación luminosa de energía $E = h\nu$, en el cual se rompe un enlace en el semiconductor (a), para dar lugar a un electrón en la banda de conducción y un hueco en la banda de valencia (b).

En estos casos por cada fotón absorbido se rompe un enlace y se crea un par electrón-hueco. También se puede crear un exceso de portadores sobre el valor de equilibrio cuando se ponen en contacto dos semiconductores de signo opuesto. Este es el caso de la unión p-n estudiado en el capítulo III, en el cual existe un trasvase de portadores de un semiconductor a otro. A todos estos procesos que originan un exceso de portadores (Δn , Δp) sobre los valores de equilibrio (n_0 , p_0) se les denomina *procesos de inyección*. Los procesos de inyección son muy comunes en el funcionamiento de los dispositivos semiconductores. Frecuentemente, en los casos denominados de *baja inyección*, el exceso de portadores solamente modifica la concentración de portadores minoritarios mientras que la de los mayoritarios no se altera apreciablemente. Así, cuando se trata de un semiconductor tipo n esta condición implica que $\Delta n \ll n_0$ y $\Delta p \gg p_0$, con $\Delta n = \Delta p$. En otros casos de *alta inyección* se modifican profundamente ambas concentraciones. Esto ocurre, por ejemplo, en algunos dispositivos electrónicos tales como los diodos láser que veremos más adelante.

2.6.1. Tiempo de vida de los portadores

Para entender mejor el comportamiento de los portadores en los procesos de inyección consideremos el caso de un semiconductor tipo n formado por una placa de espesor pequeño, con objeto de que la radiación pueda penetrar en todo su interior sin sufrir apenas atenuación. Cuando el semiconductor se ilumina durante un corto espacio de tiempo con radiación electromagnética de energía suficiente se origina en este tiempo un exceso de portadores, Δn e Δp , en las bandas conducción y valencia, respectivamente (fig.2.14a). Supondremos además que se trata de un proceso de baja inyección en el cual la concentración de portadores mayoritarios prácticamente no se altera mientras que la de minoritarios aumenta sensiblemente ($\Delta p \gg p_0$). En el instante $t = 0$, cuando cesa la irradiación, el sistema tratará de volver al equilibrio compensando el exceso de huecos mediante procesos de recombinación. En estas circunstancias, podemos establecer, en primera aproximación, que la disminución con el tiempo de la concentración de huecos en cualquier instante, $p(t)$, es proporcional al exceso de huecos, Δp , en ese instante, según una ecuación del tipo:

$$\frac{\partial p(t)}{\partial t} \Big|_{\text{recomb}} = - \frac{\Delta p}{\tau_h} \quad [2.40]$$

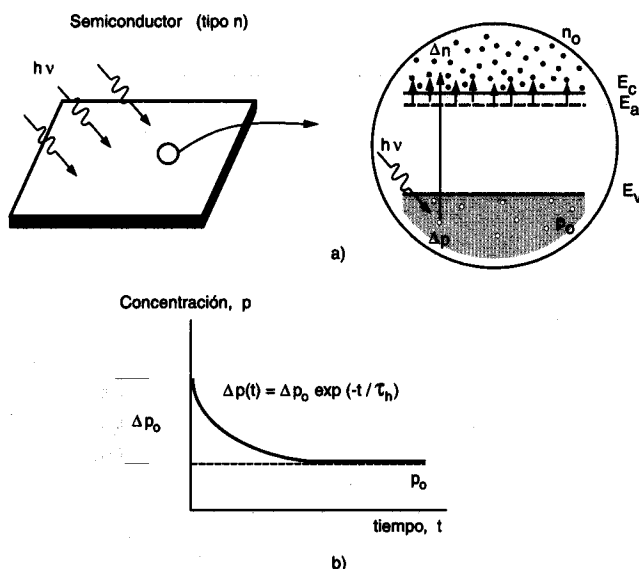


Fig.2.14. a) Inyección de portadores mediante excitación con radiación electromagnética de un semiconductor tipo n. b) Variación de la concentración de huecos con el tiempo cuando cesa la excitación.

con una ecuación similar para semiconductores tipo p. La constante de proporcionalidad que aparece en el denominador del segundo miembro, τ_h , recibe el nombre de *tiempo de vida* de los portadores minoritarios, huecos en este caso. Se puede demostrar en un análisis más riguroso que este tiempo de vida coincide con el tiempo medio que sobreviven los portadores después del proceso de inyección hasta que se alcanza el equilibrio por recombinación (es obvio que no todos los portadores duran el mismo tiempo y que unos desaparecen antes que otros).

En el ejemplo considerado, para cualquier instante de tiempo se cumplirá la ecuación:

$$p(t) = p_o + \Delta p(t) \quad [2.41]$$

con p_o constante. Por tanto, en la ec. [2.40] podemos sustituir la derivada $\partial p / \partial t$ por dp / dt , resultando:

$$\frac{d\Delta p}{dt} = - \frac{\Delta p}{\tau_h} \quad [2.42]$$

La solución de esta ecuación conduce a una disminución exponencial del exceso de huecos, con una constante de tiempo igual al tiempo de vida de los portadores, es decir:

$$\Delta p = \Delta p_o \exp (- t / \tau_h) \quad [2.43]$$

siendo Δp_o el exceso de huecos inyectados en el instante $t = 0$, (fig.2.14b).

Los tiempos de vida de los portadores minoritarios, τ_e ó τ_h , dependen mucho de las características del semiconductor y en particular de la concentración de átomos de determinadas impurezas, N_i , que se añaden intencionadamente para actuar como *centros o trampas de recombinación*. Cuanto mayor sea N_i menor es el valor de τ_e ó τ_h . Esto ocurre en algunos semiconductores (los de “gap” indirecto) en los que los procesos de recombinación directa son muy improbables ya que se requiere la interacción de dos partículas (un electrón y un hueco) moviéndose con momentos aproximadamente iguales pero de sentido opuesto. Sin embargo, la presencia de trampas de recombinación permite que una partícula sea capturada en esa posición permaneciendo en ella durante un cierto tiempo hasta que llega otra de signo opuesto y se efectúa la recombinación. Los tiempos de vida media, que son del orden de 10^{-6} s ó incluso menores cuando se añaden trampas, juegan un papel fundamental en el tiempo de respuesta de los dispositivos semiconductores (véase apartado 3.5.3).

2.6.2. Longitud de difusión

En las secciones anteriores hemos considerado separadamente diferentes procesos que pueden ocurrir en el interior de un semiconductor, tales como los de conducción o difusión y los de inyección (mediante generación y posterior recombinación) de portadores. Sin embargo es preciso señalar que todos o algunos de ellos pueden actuar conjuntamente para dar lugar a nuevos efectos. Así, por ejemplo, cuando se efectúa una inyección de portadores en una región localizada de un semiconductor se origina un aumento muy apreciable de la concentración de portadores en ese punto, sobre todo de los minoritarios, por lo que es de esperar un desplazamiento de los minoritarios hacia otras regiones del semiconductor por difusión. Sin embargo, la distancia recorrida por los portadores durante la difusión tiene una longitud limitada, ya que durante su recorrido están también sujetos a procesos de recombinación.

Consideremos el caso concreto de una barra semiconductor de tipo n en la que se inyectan portadores en un extremo mediante una iluminación constante (fig. 2.15a). Como consecuencia de ello se origina un exceso permanente de minoritarios en este extremo que, lógicamente, tienden a difundirse hacia el extremo opuesto. Sin embargo, en cualquier punto de la barra, el exceso de portadores tiende también a desaparecer mediante procesos de

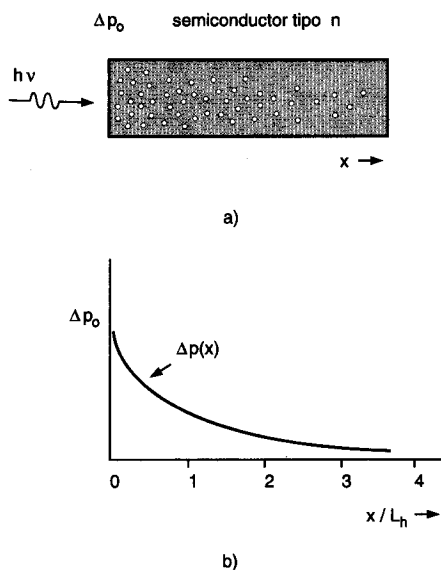


Fig.2.15. a) Esquema de la inyección de portadores en el extremo de una barra semiconductor, tipo n (mediante excitación por radiación luminosa). b) Variación del exceso de portadores minoritarios a lo largo de la barra.

recombinación. Tendremos, pues, una situación en la que los portadores generados en el extremo izquierdo recorren por difusión una cierta distancia hasta que desaparecen. Por tanto, la concentración total de portadores minoritarios va disminuyendo a lo largo del eje de la barra a medida que aumenta la distancia al origen.

Nos podemos plantear cómo varía la concentración a lo largo de la barra y la distancia recorrida por los portadores minoritarios (huecos) hasta que desaparecen por recombinación. Supondremos el caso monodimensional y condiciones de baja inyección. Para una sección elemental de la barra de anchura Δx , la variación de la concentración de huecos debido a los procesos de recombinación da lugar a una variación del flujo de huecos a través de esa sección. Este cambio en la concentración de huecos se puede expresar a través de la conocida *ecuación de continuidad*, dada en este caso por:

$$\frac{\partial J_h}{\partial x} = q \frac{\partial p}{\partial t} \quad [2.44]$$

En esta ecuación, el primer miembro refleja la variación del flujo de carga (por unidad de volumen) en el volumen elemental considerado. Este flujo de carga es debido a la corriente de difusión de los huecos (minoritarios), J_h , con valor dado por la ec. [2.33]. Tendremos por tanto de la ecuación anterior:

$$q D_h \frac{\partial^2 p}{\partial x^2} = q \frac{\partial p}{\partial t} \quad [2.45]$$

Al igual que en el ejemplo de la sec. 2.6.1 podemos expresar la variación de la concentración de huecos minoritarios mediante una ecuación similar a la ec. [2.41], es decir:

$$p(x,t) = p_o + \Delta p(x,t) \quad [2.46]$$

siendo $p(x,t)$ el exceso de huecos originados en un punto x de la barra, en el instante t y p_o la concentración de equilibrio (constante en todo el interior de la barra). La ec. [2.45] resulta entonces:

$$- D_h \frac{\partial^2 \Delta p}{\partial x^2} = \frac{\partial \Delta p}{\partial t}$$

o bien, de acuerdo con la ec. [2.42]:

$$D_h \frac{\partial^2 \Delta p}{\partial x^2} = \frac{\Delta p}{\tau_h} \quad [2.47]$$

Para resolver esta ecuación se puede establecer que la generación de pares electrón-hueco por la acción continuada de la radiación electromagnética da lugar a un exceso permanente de portadores en el origen de la barra, Δp_0 . Esta condición, junto con la que expresa que para distancias infinitas el exceso de portadores debe ser cero, es decir $\Delta p(\infty) = 0$, permite alcanzar la siguiente solución:

$$\Delta p(x) = \Delta p_0 \exp(-x/L_h) \quad [2.48]$$

siendo:

$$L_h = (D_h \tau_h)^{1/2} \quad [2.49]$$

la denominada *longitud de difusión* de los huecos. En la fig. 2.15b se ha representado la variación de $\Delta p(x)$ a lo largo de la barra. Nótese que L_h representa la longitud característica en la caída exponencial de la función $\Delta p(x)$. Se puede demostrar que esta longitud coincide con la distancia media recorrida por los minoritarios en el proceso de difusión, antes de ser aniquilados. Para el caso de un semiconductor tipo p, tendríamos igualmente para los electrones minoritarios:

$$L_e = (D_e \tau_e)^{1/2} \quad [2.50]$$

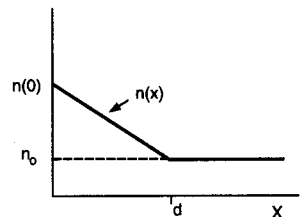
En semiconductores tipo n a la temperatura ambiente, con $\tau_h = 10^{-5}$ s, resulta para la longitud de difusión de los huecos $L_h = 3.5 \times 10^{-3}$ cm. Como veremos más adelante, las ecuaciones anteriores son muy útiles en el estudio de la unión p - n, cuando un exceso de portadores se inyecta desde un semiconductor de un tipo a otro de tipo contrario.

CUESTIONES Y PROBLEMAS

- 2.1 Con ayuda de las ecs. [1.8] y [2.2] calcular la concentración de electrones y huecos para un semiconductor tipo n a la temperatura ambiente con N_d impurezas donadoras por cm^3 parcialmente compensado con N_a impurezas aceptoras por cm^3 . Resolver también el caso opuesto cuando el semiconductor es tipo p.
- 2.2 Sabiendo que la densidad del Si es de 2.23 g/cm^3 , calcular su concentración de átomos por cm^3 . ¿Cuántos de estos átomos se encuentran ionizados a la temperatura ambiente (300 K)? ¿Cuánto vale la resistividad? ($P.A._{\text{Si}} = 28.09$, $E_g = 1.1 \text{ eV}$, $N_c = 2.8 \times 10^{19} \text{ cm}^{-3}$, $N_v = 1.04 \times 10^{19} \text{ cm}^{-3}$, $\mu_e = 0.135 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$, $\mu_h = 0.048 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$).
- 2.3 Calcular el valor de la densidad efectiva de estados para el silicio a 100 y 300 K a partir de las expresiones [2.14] y [2.15] utilizando para las masas efectivas de los electrones

y huecos los valores $m_e^* = 1.10 m_0$ y $m_h^* = 0.59 m_0$. ¿Hasta qué punto es razonable la hipótesis de que E_i está situado en el medio de la banda prohibida?

- 2.4** A partir de los datos de la figura 2.6 calcular el valor de la energía de la banda prohibida para el Si, Ge y GaAs.
- 2.5** El germanio puro tiene una banda prohibida de 0.7 eV. ¿Cuál es la probabilidad de ocupación de un estado situado en el fondo de la banda de conducción a 0 K, 300 K y 800 K?
- 2.6** Representar gráficamente la variación del nivel de Fermi respecto del nivel intrínseco de un semiconductor en función de la concentración de impurezas donadoras, N_d o aceptoras, N_a . Calcular el valor límite de N_d y N_a para el cual el semiconductor es no degenerado.
- 2.7** Un semiconductor de Si está dopado con 10^{16} átomos por cm^3 de As. Calcular la concentración de portadores y la posición del nivel de Fermi a 300 K. ¿A qué temperatura el Si puro tendría el mismo número de portadores intrínsecos? (suponer $N_c = N_v = 1 \times 10^{19} \text{ cm}^{-3}$, $E_g = 1.1 \text{ eV}$).
- 2.8** Un semiconductor de Si está dopado con $2 \times 10^{14} \text{ cm}^{-3}$ átomos de As y $3 \times 10^{14} \text{ cm}^{-3}$ de B. Suponiendo que la densidad efectiva de estados no varía con la temperatura y que la movilidad de los electrones es el doble a la de los huecos, determinar el número de portadores y el tipo de conducción a las temperaturas de 300 K y de 600 K. (datos: $E_g = 1.1 \text{ eV}$ y $n_{i(300)} = 1.5 \times 10^{10} \text{ cm}^{-3}$).
- 2.9** Discutir cómo varía cualitativamente la conductividad de un semiconductor tipo n cuando se le añaden impurezas aceptoras en cantidades crecientes.
- 2.10** Una barra semiconductor está dopada con N_{d1} y N_{d2} impurezas por cm^3 en los puntos de coordenadas x_1 y x_2 , respectivamente. Calcular la diferencia de energía del nivel de Fermi en ambos puntos antes de iniciarse el movimiento de portadores por difusión y después de alcanzarse el equilibrio.
- 2.11** En un semiconductor, la variación de la concentración de electrones a lo largo del eje x viene dada por la curva de la figura, con $n(x) = n(0) - kx$ para $0 \leq x \leq d$ y $n(x) = n_0$ para $x > d$: a) determinar la variación de la densidad de corriente, $J_e(x)$, a lo largo del eje x. b) Si el semiconductor se encuentra en equilibrio, ¿cuál es la variación del campo eléctrico generado a lo largo del eje x? c) Determinar la diferencia de potencial entre $x = 0$ y $x = d$. (suponer $n(0) / n_0 = 10^3$).



CAPITULO III

DIODOS SEMICONDUCTORES: UNION P-N

Uno de los ejemplos más característicos de los fenómenos de difusión de portadores, estudiado en el capítulo anterior, lo constituye quizás la unión de dos semiconductores cada uno de ellos con portadores mayoritarios de signo opuesto. Este conjunto denominado diodo representa el elemento más básico entre los dispositivos semiconductores de estado sólido. Los diodos ofrecen diferente resistencia al paso de la corriente según sea la polaridad de la tensión externa aplicada en sus extremos. Este efecto está determinado fundamentalmente por la carga espacial que se origina en la interfase de la unión de los dos semiconductores (unión p-n). La interfase de la unión tiene un espesor muy pequeño comparado con el resto del diodo. Sin embargo, como veremos más adelante, aún siendo pequeña esta región es la que ejerce todo el control de las propiedades del dispositivo. El estudio de los diodos de unión p-n, así como el de otros dispositivos rectificadores tales como la unión metal-semiconductor, constituye el aspecto más fundamental de este capítulo y del siguiente.

3.1. LA UNION P-N

La unión de dos semiconductores extrínsecos, uno de ellos dopado con impurezas de tipo donador y otro con impurezas de tipo aceptor, constituye lo que se denomina la unión p-n. Normalmente los semiconductores están formados por el mismo material, en el que se introducen, mediante procesos de difusión que serán descritos en el cap. XIII, impurezas de uno y otro signo en dos regiones del semiconductor adyacentes entre sí. En la fig. 3.1 se muestra un

esquema de la sección transversal de una oblea de silicio en la que se ha preparado una unión p-n para formar un diodo. Para conseguir esta unión, se parte de una oblea del material semiconductor, tipo n por ejemplo, y en una región de la oblea se añade un dopaje con impurezas de tipo opuesto, tipo p en este caso, en concentración suficiente para invertir el signo de los portadores en esa región. Finalmente, sobre la región dopada y en la parte inferior de la oblea se deposita una capa conductora con objeto de hacer los contactos eléctricos hacia el exterior.

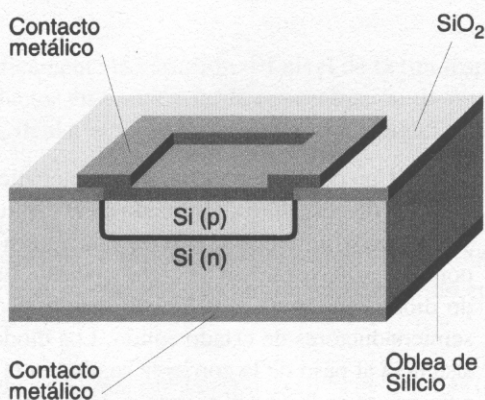


Fig. 3.1. Esquema de un diodo semiconductor formado por la unión de dos semiconductores de signo opuesto (unión p-n).

En lo que sigue, haremos una descripción cualitativa del comportamiento de la unión en el caso ideal, en el cual se supone que cada lado de la unión está formado por impurezas de un solo tipo con una distribución completamente homogénea y que el semiconductor se encuentra en equilibrio termodinámico.

3.1.1. Comportamiento de la unión p-n sin polarización externa

Supongamos primero que los dos materiales de tipo p y n se encuentran separados. En la fig. 3.2a se da el diagrama de bandas de energía correspondiente a cada uno de ellos. La figura incluye también un esquema de la distribución de impurezas y de los portadores de carga en el interior de ambos materiales semiconductores. Recordemos que a temperatura ambiente el nivel de Fermi, E_F , tiene un valor próximo a la energía E_c de la banda de conducción en el semiconductor tipo n, mientras que se acerca al valor E_v en el semiconductor tipo p. Sobre el diagrama de bandas de energía se ha trazado de forma cualitativa las curvas de distribución de la energía de los portadores en cada uno de los semiconductores.

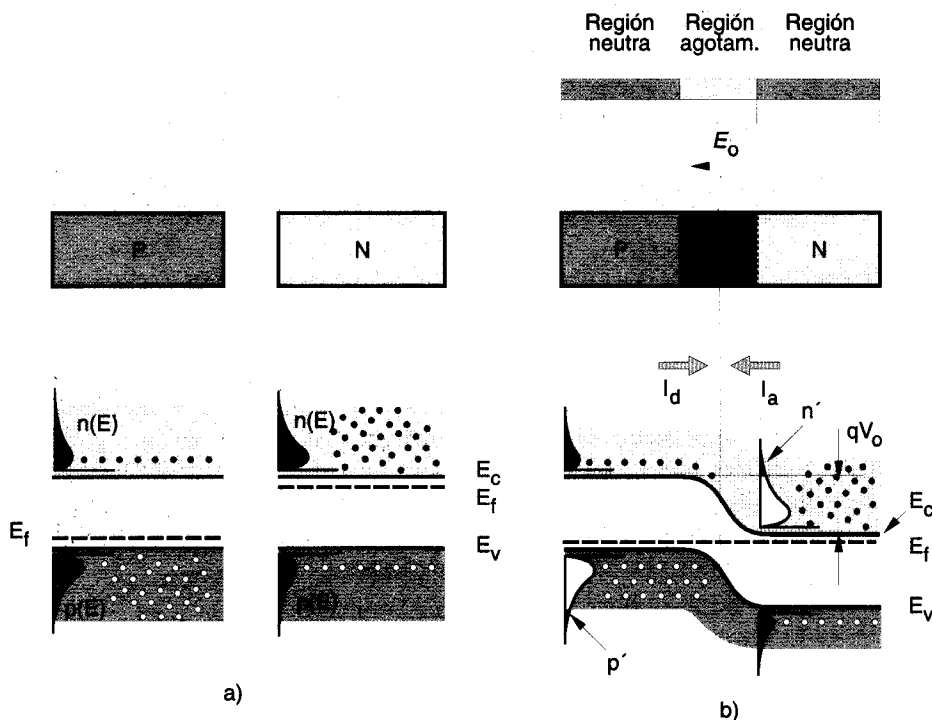


Fig.3.2. a) Esquema de las bandas de energía y de la función de distribución de electrones y huecos en dos semiconductores extrínsecos, tipos p y n, cuando están separados. b) Esquema de la carga espacial y de las bandas de energía una vez que los semiconductores se juntan formando la unión p-n.

Cuando los semiconductores están en contacto (fig.3.2b), la diferencia en la concentración de portadores a uno y otro lado de la unión da lugar a que existan corrientes de difusión para alcanzar el equilibrio termodinámico. En el lado n de la unión tendremos una difusión de electrones mayoritarios hacia el lado p moviéndose en la banda de conducción, y en el lado p tendremos una corriente de huecos, también mayoritarios, que cruzan en dirección opuesta hacia el lado n por la banda de valencia.

Después de cruzar la unión, los electrones y huecos mayoritarios se recombinan en el lado opuesto, originando una *región de carga espacial* en las proximidades de la unión (positiva en el lado n y negativa en el lado p). La región de carga espacial está asociada a la carga de las impurezas ionizadas que quedan sin compensar a cada lado de la unión. En esta región

prácticamente no existe carga libre y por esta razón se la denomina también *región de agotamiento*. Según vimos en el capítulo anterior, el proceso de difusión continúa hasta que el campo eléctrico E_0 debido a la carga espacial se opone al movimiento de carga debido a la difusión. La presencia de este campo a su vez da lugar a una diferencia de potencial V_0 entre uno y otro lado, de forma que el diagrama de bandas de energía del conjunto se modifica según se esquematiza en la parte inferior de la fig. 3.2b, manteniéndose en todo caso el nivel de Fermi constante. Se observa que existe una disminución de las bandas de energía al pasar del lado p al n, representada por el valor qV_0 , la cual viene dada por la diferencia de energías de los niveles de Fermi cuando los semiconductores estaban separados.

En conjunto se puede decir que una vez alcanzado el equilibrio los electrones del lado n y los huecos del lado p quedan confinados en las denominadas *regiones neutras* de cada lado de la unión, separados por la región de carga espacial en la cual existe una barrera de potencial de altura qV_0 . Solamente una pequeña fracción de los portadores mayoritarios de uno y otro lado tienen energía superior a la barrera (área sombreada en las curvas de distribución de la fig. 3.2b), y son los que pueden pasar por difusión al lado opuesto de la unión. Sin embargo, en el equilibrio este movimiento de carga por difusión hacia un lado, corriente I_d , está compensado siempre por una corriente de arrastre, I_a , en sentido opuesto, debida al campo eléctrico E_0 originado en las proximidades de la unión (nótese que el movimiento de los electrones se realiza en sentido contrario al señalado para las corrientes).

3.1.2. La unión p-n polarizada con un voltaje externo

Cuando se aplica un voltaje externo, V , la situación de equilibrio en la unión p-n queda modificada. **La región de agotamiento, debido a la ausencia de portadores en ella, ofrece una resistencia muy elevada frente a las regiones neutras de tipo p ó n.** Por esta razón, todo el potencial V cae prácticamente en la región de agotamiento dando lugar con ello a un desplazamiento de la energía de las bandas igual a qV . Para tensiones positivas aplicadas en el lado p y negativas en el lado n (*polarización en directo*), tanto los huecos del lado p como los electrones del lado n adquieren mayor energía potencial, por lo que la altura de la barrera se reduce en qV , y al mismo tiempo disminuye la anchura de la región de carga espacial. La distribución en energía de los electrones y de los huecos también se modifica, dando lugar a que una mayor fracción de electrones y huecos mayoritarios (área sombreada en las curvas de distribución de la fig. 3.3) pueda pasar al lado opuesto por difusión, mientras que el movimiento de arrastre de los portadores minoritarios en sentido opuesto, debido al campo eléctrico E presente en la unión, queda prácticamente inalterado. En estas circunstancias, la corriente de difusión de los portadores mayoritarios, I_d , aumenta en relación a la corriente de arrastre, I_a . Se obtiene así una corriente grande de electrones hacia el electrodo positivo y de huecos hacia el electrodo negativo (fig. 3.3a).

Si la tensión V aplicada tiene signo opuesto, es decir el lado n positivo frente al lado p (*polarización en inverso*), ocurre un fenómeno similar, aunque en este caso se produce un

desplazamiento de las bandas en sentido opuesto al anterior, disminuyendo la energía de los electrones en el lado n y la de los huecos en el lado p. La altura de la barrera aumenta en una cantidad igual a qV y la región de carga espacial se ensancha. Ahora existe una menor fracción de electrones y huecos mayoritarios con energía suficiente para cruzar la barrera, por lo que el desplazamiento de los portadores mayoritarios por difusión se reduce frente al movimiento de los portadores minoritarios por arrastre del campo eléctrico. Dado que la concentración de portadores minoritarios es siempre muy baja, la corriente neta a través de la unión es en estas condiciones muy baja. Además, el sentido de esta corriente es opuesto a la corriente de difusión de los portadores mayoritarios. (fig. 3.3b).

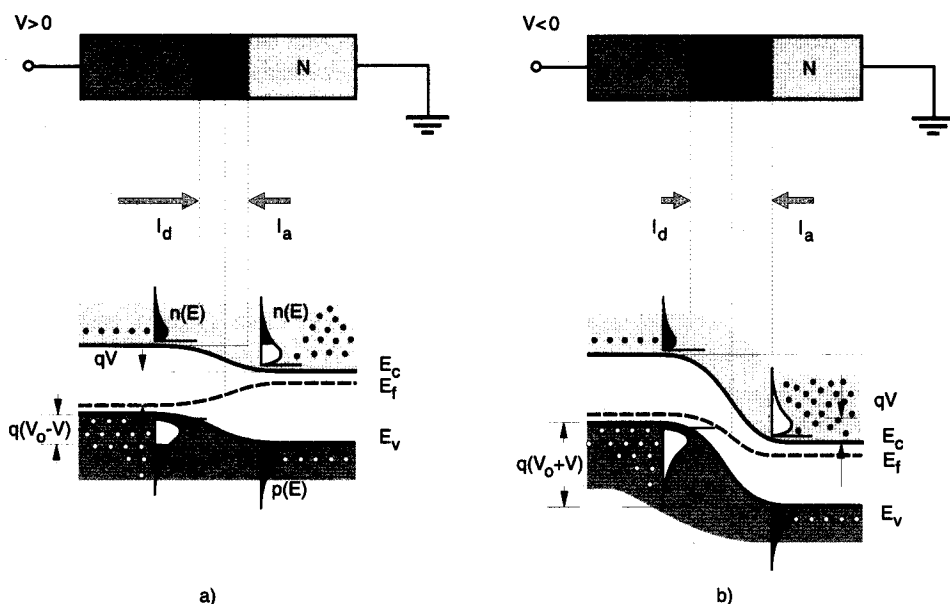


Fig.3.3. Representación de las bandas de energía y de la distribución de electrones y huecos cuando la unión p-n se polariza, a) en directo y b) en inverso.

Es importante resaltar que cuando existe un voltaje externo, V , el nivel de Fermi en el sistema ya no es constante. Así, cuando se aplica una polarización en directo el nivel de Fermi en el lado p estará situado a menor energía que en el lado n, siendo la diferencia de energías igual a qV . La altura de la barrera de potencial para los electrones del lado n y huecos lado p se reduce entonces en esa cantidad, pasando a ser $q(V_0 - V)$. En polarización en inversa ocurre lo contrario, el nivel de Fermi del lado p está a mayor energía que el del lado n y la altura de la

barrera para los portadores mayoritarios se incrementa en qV , de forma que la altura total de barrera es $q(V_o + V)$.

3.2. VARIACION DEL POTENCIAL EN LA REGION DE CARGA ESPACIAL

Como hemos visto, la región de carga espacial se extiende desde el centro hasta una cierta distancia a cada lado de la unión. La mayor parte de la carga contenida en ella, en forma de doble capa (positiva y negativa), es **carga fija** procedente de las impurezas ionizadas sin compensar. El origen de esta descompensación es doble: por un lado está la disminución de los portadores mayoritarios en las regiones p y n como consecuencia del trasvase de huecos y electrones a uno y otro lado de la unión. Por otro lado, los portadores mayoritarios cuando pasan al lado opuesto se recombinan en una distancia igual a su longitud de difusión. La recombinación produce por tanto una disminución adicional de los portadores en esta zona, y hace que la región de carga espacial esté prácticamente vacía de portadores libres. Esta carga fija es la responsable del campo eléctrico, E_o , presente en la región de carga espacial y del correspondiente potencial de contacto, V_o , cuando la unión se encuentra en equilibrio térmico, sin ningún potencial externo aplicado.

Como punto de partida, podemos calcular el valor de la caída de potencial asociada a la región de carga espacial utilizando el principio de igualdad del nivel de Fermi cuando el semiconductor alcanza el equilibrio. La energía asociada a la caída de potencial, qV_o , será igual a la diferencia de energía entre los niveles de Fermi en cada lado de la unión en el estado inicial, es decir:

$$qV_o = (E_f)_n - (E_f)_p$$

siendo $(E_f)_n$ y $(E_f)_p$ la energía del nivel de Fermi del lado n y del lado p, respectivamente. El valor de $(E_f)_n$ y de $(E_f)_p$ se puede calcular directamente de las ecs. [2.17] y [2.18] si suponemos que a la temperatura ambiente todas las impurezas están ionizadas, con $n=N_d$ en el lado n y $p=N_a$ en el lado p. Tendremos entonces:

$$(E_f)_n = E_i + kT \ln \frac{N_d}{n_i} \quad [3.1]$$

y

$$(E_f)_p = E_i - kT \ln \frac{N_a}{n_i} \quad [3.2]$$

resultando:

$$V_o = \frac{(E_f)_n - (E_f)_p}{q} = \frac{kT}{q} \ln \frac{N_a N_d}{n_i^2} \quad [3.3]$$

Este resultado muestra que cuanto mayor sea la concentración de impurezas de los lados p y n mayor será la caída de potencial en la unión, V_o .

Es interesante conocer cómo varía el potencial, $V(x)$, a lo largo de la unión. Para ello, podemos suponer que, en primera aproximación, la concentración de carga fija en la región de carga espacial coincide con la concentración de impurezas ionizadas a uno y otro lado de la unión, dada por $+qN_d$ en el lado n y $-qN_a$ en el lado p. Al mismo tiempo, la concentración de carga libre (electrones y huecos) debe mantenerse en una proporción muy baja en esta región. La función $V(x)$ se puede obtener entonces aplicando la ecuación de Poisson de la electrostática a la región de carga espacial. Para una sola dimensión tendremos:

$$\frac{\partial^2 V}{\partial x^2} = - \frac{\rho(x)}{\epsilon} \quad [3.4]$$

siendo ϵ la constante dieléctrica del semiconductor y $\rho(x)$ la densidad o concentración de carga en cada punto de la región de carga espacial.

Podemos plantearnos el caso simple en el que las concentraciones N_a y N_d son constantes en el interior del semiconductor (*unión abrupta*), y además los bordes de la región espacial terminan también con una variación abrupta a distancias x_n y $-x_p$, medidas desde el centro de la unión hacia los lados n y p, respectivamente (fig. 3.4a). Eligiendo el centro de la unión como origen de coordenadas, tendremos entonces para la densidad de carga a cada lado de la unión:

$$\rho(x) = + qN_d \quad (0 < x \leq x_n)$$

$$\rho(x) = - qN_a \quad (0 > x \geq -x_p)$$

En la fig. 3.4b se muestra un esquema de la curva de la densidad de carga en la región de carga espacial.

La ecuación de Poisson se puede descomponer en dos ecuaciones para cada uno de los lados, de forma que para el lado n de la región de carga espacial tendremos:

$$\frac{d^2 V}{dx^2} = - \frac{q N_d}{\epsilon} \quad [3.5]$$

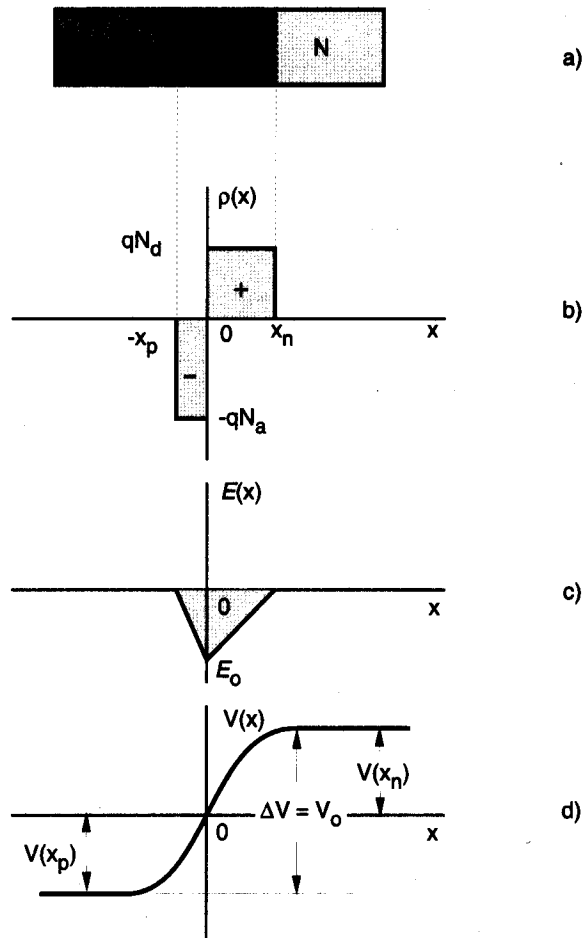


Fig.3.4. Esquema del espesor (a), de la distribución de carga (b), del campo eléctrico (c) y del potencial (d) en la región de carga espacial en un diodo sin polarización externa.

con una ecuación similar para el lado p. La integración de esta ecuación, con la condición $E = -dV/dx = 0$ para $x = x_n$, ya que fuera de la región de carga espacial el campo eléctrico es cero (región neutra), conduce a:

$$E(x) = - \frac{dV}{dx} = \frac{q N_d}{\epsilon} (x - x_n) \quad [3.6]$$

para valores de x en el intervalo $0 < x \leq x_n$. Análogamente para el potencial, eligiendo como potencial de referencia el origen de coordenadas, esto es $V(x) = 0$ para $x = 0$:

$$V(x) = -\frac{q N_d}{2\epsilon}(x^2 - 2x_n x) \quad [3.7]$$

para $0 < x \leq x_n$. Una variación similar para $E(x)$ y $V(x)$ se obtendría para el lado p de la región de carga espacial. En las figuras 3.4c y 3.4d se da una representación esquemática del campo eléctrico y del potencial a cada lado de la unión, conforme a las ecuaciones [3.6] y [3.7] y las equivalentes para el lado p. Nótese que el campo eléctrico es siempre negativo y tiene una variación lineal dentro de la región de carga espacial disminuyendo en valor absoluto desde el centro de la unión hasta los bordes con las zonas neutras. En el centro de la unión ($x=0$) es donde el campo eléctrico toma el máximo valor absoluto debido a la proximidad de las cargas positivas y negativas.

Supuesto conocidos la anchura de cada una de las regiones de carga espacial podemos calcular las caídas de voltaje, $V(x_n)$ y $V(-x_p)$, asociadas a ellas. Así, si sustituimos x por x_n en la ec. [3.7] resulta:

$$V(x_n) = \frac{q N_d x_n^2}{2\epsilon} \quad [3.8]$$

Análogamente, para el lado p tendríamos:

$$V(-x_p) = -\frac{q N_a x_p^2}{2\epsilon} \quad [3.9]$$

Por tanto, el incremento total del voltaje, $V(x_n) - V(-x_p)$, al pasar del lado n al lado p, cuyo valor debe coincidir con el potencial de contacto asociado a la unión, V_o , vendrá dado por:

$$V_o = \frac{q}{2e} (N_d x_n^2 + N_a x_p^2) \quad [3.10]$$

La ecuación anterior relaciona V_o con el espesor de cada una de las capas de la región de carga espacial. Dado que V_o puede ser conocido a través de la ec. [3.3], es posible obtener separadamente los valores de x_n y x_p si utilizamos junto a la ecuación anterior la condición de neutralidad de carga:

$$q N_d x_n = q N_a x_p \quad [3.11]$$

la cual expresa que la carga por unidad de superficie, Q' , contenida en cada lado de la unión en la región de carga espacial debe ser la misma. Esta ecuación pone además de manifiesto que

las anchuras de cada lado de la región de carga espacial son inversamente proporcionales a los dopajes respectivos de cada región. Tomando como incógnitas x_n y x_p , las ecs. [3.10] y [3.11] forman un sistema de ecuaciones cuya solución es:

$$x_n = \left[\frac{2\epsilon V_o}{q N_d} \left(\frac{N_a}{N_d + N_a} \right) \right]^{1/2} \quad [3.12]$$

y

$$x_p = \left[\frac{2\epsilon V_o}{q N_a} \left(\frac{N_d}{N_d + N_a} \right) \right]^{1/2} \quad [3.13]$$

Llamando x_o al espesor total de la región de carga espacial, es decir, $x_o = x_n + x_p$, resulta finalmente:

$$x_o = \left[\frac{2\epsilon V_o}{q} \left(\frac{1}{N_d} + \frac{1}{N_a} \right) \right]^{1/2} \quad [3.14]$$

A menudo, la unión se prepara con uno de los lados mucho más dopado que el otro. Si por ejemplo $N_a \gg N_d$, *unión abrupta $p^+ - n$* , resulta $x_n \gg x_p$, y $x_o \approx x_n$, con x_o dado por:

$$x_o = \left(\frac{2\epsilon V_o}{q N_d} \right)^{1/2} \quad [3.15]$$

Al aplicar al diodo una tensión externa, V , existe una variación de la carga acumulada a uno y otro lado de la región de carga espacial, por lo que la anchura de esta región también varía. Podemos suponer en primera aproximación que el voltaje externo V se superpone con signo negativo al que ya existe en la barrera, V_o , por lo que la relación entre el espesor total resultante, x , y el voltaje aplicado será similar a la ec. [3.14], cambiando V_o por $V_o - V$ y x_o por x . Resulta así:

$$x = \left[\frac{2\epsilon (V_o - V)}{q} \left(\frac{1}{N_d} + \frac{1}{N_a} \right) \right]^{1/2} \quad [3.16]$$

Debido al signo de V , esta relación muestra que a medida que crece el voltaje en polarización directa (con un valor máximo inferior a V_o) el espesor total de la región de carga espacial disminuye, según hemos visto en el apartado anterior (véase fig. 3.3). En cambio, para polarización inversa ocurre lo contrario y la región de carga espacial se ensancha a medida que aumenta V en sentido negativo. Además, para un voltaje dado el espesor de esta región disminuye al aumentar la concentración de impurezas en el semiconductor. Esta propiedad

tiene gran importancia en la preparación de diodos en los que interesa por ejemplo que la región de carga espacial sea lo más estrecha posible, como es el caso de los diodos Zener (véase sec. 3.3.4).

3.3. CALCULO DE LA CORRIENTE A TRAVES DE LA UNION P-N

Nos podemos plantear ahora el cálculo de la corriente a través de la unión p-n cuando se polariza con un voltaje, V , determinado. Veamos primero cuál es la concentración de portadores a uno y otro lado de la unión.

3.3.1. Concentración de portadores cuando no hay voltaje aplicado ($V=0$)

Como situación de partida, estudiaremos la unión sin polarizar en estado de equilibrio. De acuerdo con la descripción del apartado anterior, como consecuencia del trasvase de electrones del lado n y de huecos del lado p en sentido opuesto, se establece en la unión una diferencia de potencial, V_o , cuyo valor viene determinado por las concentraciones de portadores de uno y otro signo a cada lado de la unión. Según la ecuación [3.3], V_o viene dado por:

$$V_o = \frac{kT}{q} \ln \frac{p_{po} n_{no}}{n_i^2} = \frac{kT}{q} \ln \frac{n_{no}}{n_{po}} \quad [3.17]$$

donde se ha sustituido N_a y N_d por las concentraciones de equilibrio de los portadores mayoritarios, p_{po} y n_{no} , es decir, de huecos en el lado p y electrones en el lado n, respectivamente. En la última igualdad se ha hecho uso de la ley de acción de masas, $p_{po} n_{po} = n_i^2$, para la concentración de electrones y huecos en la zona neutra del lado p. Supondremos que las zonas neutras de los lados p y n del diodo se extienden desde los extremos del diodo hasta distancias x_p y x_n de la unión, respectivamente.

La ecuación [3.17] permite relacionar la concentración de electrones (portadores mayoritarios) en la zona neutra de carácter n con la que existe en la zona neutra de carácter p (portadores minoritarios), de forma que esta última concentración se puede escribir como:

$$n_{po} = n_{no} \exp(-qV_o/kT) \quad [3.18]$$

y del mismo modo, la concentración de huecos (minoritarios) en el lado n en función de los huecos en el lado p (mayoritarios) vendrá dada por:

$$p_{no} = p_{po} \exp(-qV_o/kT) \quad [3.19]$$

Hemos visto en la sección 3.1.1 que solamente los electrones del lado n (y los huecos del lado p) con energía E superior a la barrera ($E > qV_o$) son capaces de cruzar la unión e intercambiarse con los electrones (y huecos) minoritarios del lado opuesto. Si denominamos n'_{no} y p'_{po} a las fracciones de electrones en el lado n, y de huecos en el lado p, respectivamente, con energía $E > qV_o$ (área rayada en las curvas de distribución de portadores de la fig. 3.2b), los valores n'_{no} y p'_{po} deben ser en cada caso aproximadamente iguales a las concentraciones totales de electrones y huecos minoritarios en el lado opuesto. Esto quiere decir que cuando la unión se encuentra en equilibrio podemos escribir:

$$n'_{no} \approx n_{po} = n_{no} \exp(-qV_o/kT) \quad [3.20]$$

$$p'_{po} \approx p_{no} = p_{po} \exp(-qV_o/kT) \quad [3.21]$$

3.3.2. Concentración de portadores con un voltaje externo aplicado ($V \neq 0$)

Cuando se aplica una tensión externa, V , en los extremos del diodo, la altura de la barrera se modifica tomando el valor ($V_o - V$) según hemos visto más arriba. Si la tensión es positiva, polarización en directo, se favorece el trasvase o *inyección* de electrones del lado n al p y de huecos en sentido opuesto. Por tanto, la concentración de electrones en el lado p, y de huecos en el lado n, en las proximidades de la unión aumenta sobre la que existía anteriormente (ecs. 3.18 y 3.19). Cuando la tensión es negativa ocurre el proceso opuesto y las concentraciones de portadores minoritarios a ambos lados de la unión disminuyen. En lo que sigue supondremos que en polarización positiva existen condiciones de baja inyección (véase apartado 2.6), lo cual ocurre siempre que $V \ll V_o$. Esta condición implica que la concentración total de portadores mayoritarios a uno y otro lado de la unión, que denominaremos n_n y p_p para los lados n y p, respectivamente, prácticamente no varía por el hecho del trasvase de portadores, mientras que la concentración de los portadores minoritarios puede verse seriamente alterada en las proximidades de la unión. Cuando las regiones neutras son suficientemente largas, podemos suponer además que a distancias alejadas de la unión las concentraciones de los portadores minoritarios no se modifican apenas. Si n_p y p_n representan las concentraciones de minoritarios en los lados p y n, respectivamente, después de aplicar un voltaje externo a la unión, según esta última hipótesis tendremos que: $n_p(\infty) \approx n_{po}$ y $p_n(\infty) \approx p_{no}$. Al mismo tiempo se cumple también para los mayoritarios, lógicamente, que $n_n = n_{no}$ y $p_p = p_{po}$.

La aplicación de una tensión externa en los extremos del diodo implica que la energía de los electrones y huecos mayoritarios aumenta o disminuye en relación a la energía de los minoritarios en la cantidad qV . Así pues, si llamamos n'_n y p'_p a las concentraciones de electrones y huecos mayoritarios a ambos lados de la unión con energía $E \geq q(V_o - V)$ (área rayada en las curvas de distribución de la fig. 3.5), podremos escribir en analogía con las ecs. [3.20] y [3.21]:

$$n_n' \approx n_n \exp [-q (V_o - V) / kT] = n_{no} \exp [-q(V_o - V) / kT] \quad [3.22]$$

$$p_p' \approx p_p \exp [-q (V_o - V) / kT] = p_{po} \exp [-q(V_o - V) / kT] \quad [3.23]$$

O bien, utilizando de nuevo las ecs. [3.20] y [3.21]:

$$n_n' \approx n_{po} \exp (qV / kT) \quad [3.24]$$

$$p_p' \approx p_{no} \exp (qV / kT) \quad [3.25]$$

Al igual que en el caso $V = 0$, estos portadores mayoritarios, es decir, los que tienen energía $E \geq q (V_o - V)$, son los que se intercambian con los minoritarios del lado opuesto produciendo un flujo de corriente positivo o negativo (según sea el signo de V) a través de la unión. Así pues, se puede establecer que la fracción de electrones y de huecos mayoritarios que cruza al otro lado de la unión viene dado por las ecuaciones anteriores. Por tanto, denominando $n_p(0)$ y $p_n(0)$ las concentraciones de electrones y huecos minoritarios justo en los bordes de la región de carga espacial, tendremos:

$$n_p(0) = n_n' = n_{po} \exp (qV / kT) \quad [3.26]$$

$$p_n(0) = p_p' = p_{no} \exp (qV / kT) \quad [3.27]$$

Este resultado muestra que, en polarización directa ($V > 0$), la concentración de minoritarios en los bordes de la región neutra aumenta sobre los valores en equilibrio, n_{po} y p_{no} , como consecuencia del trasvase de mayoritarios desde el lado opuesto de la unión. En cambio si la polarización es inversa ($V < 0$) ocurre el efecto contrario, es decir, disminuyen las concentraciones de minoritarios ya que en este caso son los minoritarios los que pasan al otro lado de la unión.

De acuerdo con las expresiones anteriores, podemos calcular el exceso (o defecto) de portadores minoritarios a ambos lados de barrera, $\Delta n_p(0)$ y $\Delta p_n(0)$, sobre los valores normales de equilibrio a distancias grandes de la unión, esto es $n_p(\infty)$ y $p_n(\infty)$, obteniéndose:

$$\Delta n_p(0) = n_p(0) - n_p(\infty) = n_{po} [\exp (qV / kT) - 1] \quad [3.28]$$

$$\Delta p_n(0) = p_n(0) - p_n(\infty) = p_{no} [\exp (qV / kT) - 1] \quad [3.29]$$

Llegados a este punto es importante subrayar que el exceso de portadores minoritarios no se mantiene constante en las respectivas zonas neutras ya que los portadores una vez que cruzan la barrera se difunden hacia el lado opuesto hasta que se recombinan. La longitud de difusión, esto es, la distancia media recorrida hasta que desaparecen por recombinación, es en general menor que la distancia hasta el electrodo opuesto. Al mismo tiempo, los electrones y

huecos que desaparecen por recombinación son compensados por nuevos portadores que proceden del lado opuesto manteniendo la continuidad de la corriente.

En este proceso continuo de inyección, desplazamiento por difusión y recombinación de portadores que tiene lugar en cada uno de los lados de la unión p-n, existe una variación de la concentración de minoritarios a lo largo de la distancia a la unión. Se puede hacer un cálculo de la variación de los minoritarios utilizando la ecuación de la continuidad, estudiada en el ca-

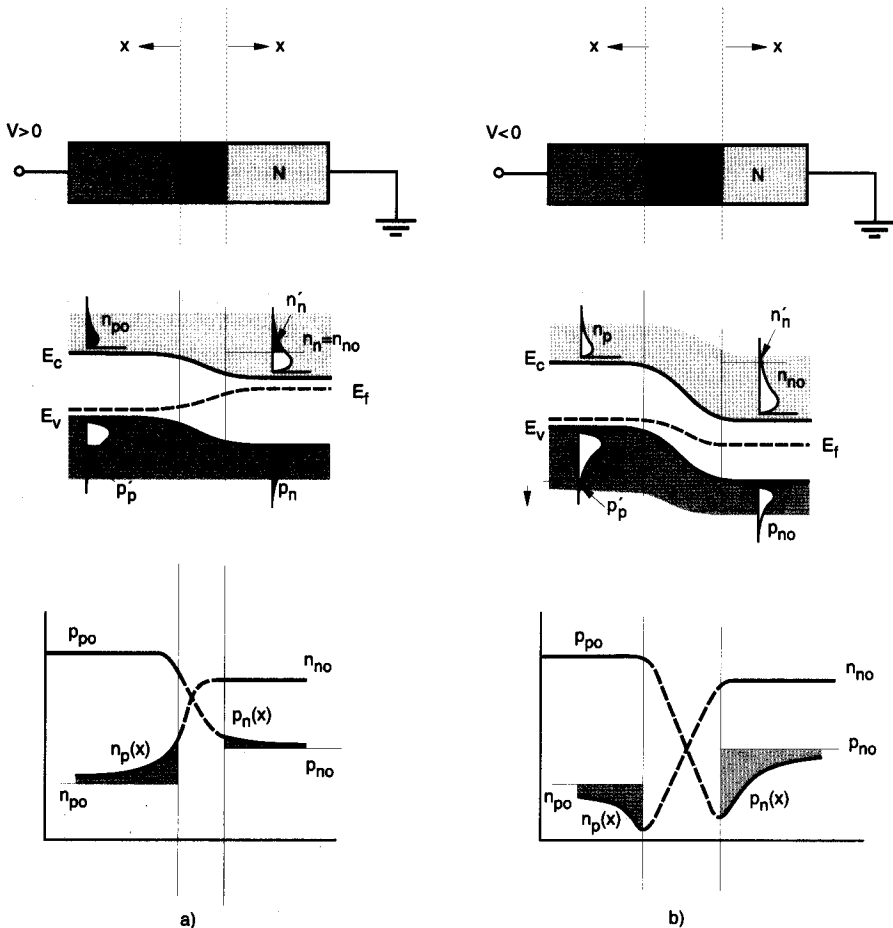


Fig.3.5. Representación de las bandas de energía y de la concentración de portadores mayoritarios y minoritarios en cada lado de las regiones neutras cuando la unión se polariza, a) en directo y b) en inverso.

pítulo anterior. Se trata ahora de una situación muy similar a la del ejemplo de la sección 2.6.2, en la cual teníamos también una inyección permanente de portadores en un extremo del semiconductor dando lugar a un exceso de minoritarios. De acuerdo con los resultados obtenidos entonces, el exceso de minoritarios sufre una disminución exponencial a medida que aumenta la distancia x a ese extremo (en este caso los bordes de la región de agotamiento). Por tanto, Δn_p y Δp_n variarán de acuerdo con las ecuaciones:

$$\Delta n_p(x) = \Delta n_p(0) \exp(-x / L_e) \quad [3.30]$$

para la zona neutra de tipo p, y:

$$\Delta p_n(x) = \Delta p_n(0) \exp(-x / L_h) \quad [3.31]$$

para la zona neutra de tipo n. En estas ecuaciones L_e y L_h representan las longitudes de difusión de los electrones y huecos minoritarios en las zonas p y n, respectivamente, y vienen dadas por las ecs. [2.49] y [2.50].

En las figs. 3.5a y 3.5b (parte inferior) se ha hecho una representación esquemática de la variación de portadores en las zonas neutras próximas a ambos lados de la región de carga espacial. Obsérvese en la zona neutra la variación exponencial de la concentración de minoritarios, $n_p(x)$ y $p_n(x)$ según las ecuaciones:

$$n_p(x) = n_{p0} + \Delta n_p(x) \quad [3.32]$$

$$p_n(x) = p_{n0} + \Delta p_n(x) \quad [3.33]$$

con $\Delta n_p(x)$ y $\Delta p_n(x)$ dados por [3.30] y [3.31].

3.3.3. Cálculo de la corriente

Hemos visto que, cuando la unión se polariza en directo, los portadores mayoritarios se desplazan hacia el lado opuesto para terminar recombinándose con portadores de signo contrario. Como consecuencia de estos procesos de recombinación, ha de existir un aporte continuo de portadores, electrones desde el lado n hacia el lado p y de huecos en sentido opuesto, con objeto de mantener constante la concentración de minoritarios en cada punto. Este aporte continuo de carga constituye de hecho la verdadera corriente a través del diodo. Se trata pues de un proceso de difusión, en el cual la corriente de difusión es proporcional a la derivada de la concentración de portadores, según vimos en el capítulo anterior (ecs. 2.33 y 2.34). Así, la densidad de corriente $J_g(0)$ justo en el límite de la zona neutra de carácter p viene dada, de

acuerdo con la ec. [3.32], por:

$$J_e(0) = q D_e \left. \frac{dn_p(x)}{dx} \right|_{x=0} = \frac{q D_e n_{po}}{L_e} \left[\exp \left(\frac{qV}{kT} \right) - 1 \right] \quad [3.34]$$

y en el punto correspondiente del lado n, la densidad de corriente de huecos, $J_h(0)$, será:

$$J_h(0) = q D_h \left. \frac{dp_n(x)}{dx} \right|_{x=0} = \frac{q D_h p_{no}}{L_h} \left[\exp \left(\frac{qV}{kT} \right) - 1 \right] \quad [3.35]$$

La corriente $J_e(0)$ es debida al movimiento de electrones hacia la izquierda, es decir, desde el lado n al lado p cuando la polarización es positiva (y en la dirección opuesta cuando es negativa). Al contrario, $J_h(0)$ es debida al movimiento de huecos siempre en sentido opuesto al de electrones. Ambas corrientes contribuyen por tanto con el mismo signo al valor total de la corriente.

Evidentemente, las corrientes de electrones y huecos varían al pasar de un punto a otro en cada una de las regiones neutras. Debido al carácter exponencial de $n_p(x)$ y $p_n(x)$ (ecs. 3.32 y 3.33), las corrientes $J_e(x)$ y $J_h(x)$ tienen también una variación exponencial a lo largo del eje x. La variación de J_e y J_h con la posición es consecuencia directa de los procesos de recombinación de los portadores que cruzan la unión, según se ha expuesto anteriormente. Sin embargo, la condición de continuidad de la corriente implica que en el caso estacionario la corriente total, $J_e(x) + J_h(x)$, en cualquier punto x del semiconductor debe ser constante.

Podemos calcular el valor de la corriente total justo en los bordes de la región de carga espacial si suponemos que en el interior de toda esta región J_e y J_h prácticamente no varían, tal como se indica en el esquema de la fig. 3.6a. Se obtiene así, a partir de las ecs. [3.34] y [3.35], la conocida *ecuación de Shockley* para el diodo:

$$J = J_e(0) + J_h(0) = J_o [\exp (qV / kT) - 1] \quad [3.36]$$

siendo:

$$J_o = \frac{q D_e n_{po}}{L_e} + \frac{q D_h p_{no}}{L_h} = q n_i^2 \left(\frac{D_e}{L_e N_a} + \frac{D_h}{L_h N_d} \right) \quad [3.37]$$

la denominada *corriente inversa de saturación*. En la última igualdad se ha hecho uso de nuevo de la ley de acción de masas para cada una de las zonas neutras: $n_{po} \cdot p_{po} = n_i^2$ y

$n_{no} \cdot p_{no} = n_i^2$, con $p_{po} = N_a$ y $n_{no} = N_d$. Como hemos visto, a menudo los diodos se forman con un dopaje muy elevado en una de las zonas. Así, por ejemplo, en el diodo de unión abrupta de tipo $p^+ - n$ se tiene $N_a \gg N_d$. En este caso se cumple que $n_{po} \ll p_{no}$ por lo que J_o está determinada fundamentalmente por la concentración de minoritarios (huecos) en el lado n . Algo similar ocurriría en el caso opuesto. Estos resultados revelan que la corriente de saturación está dominada por la concentración de minoritarios en cada lado de la unión.

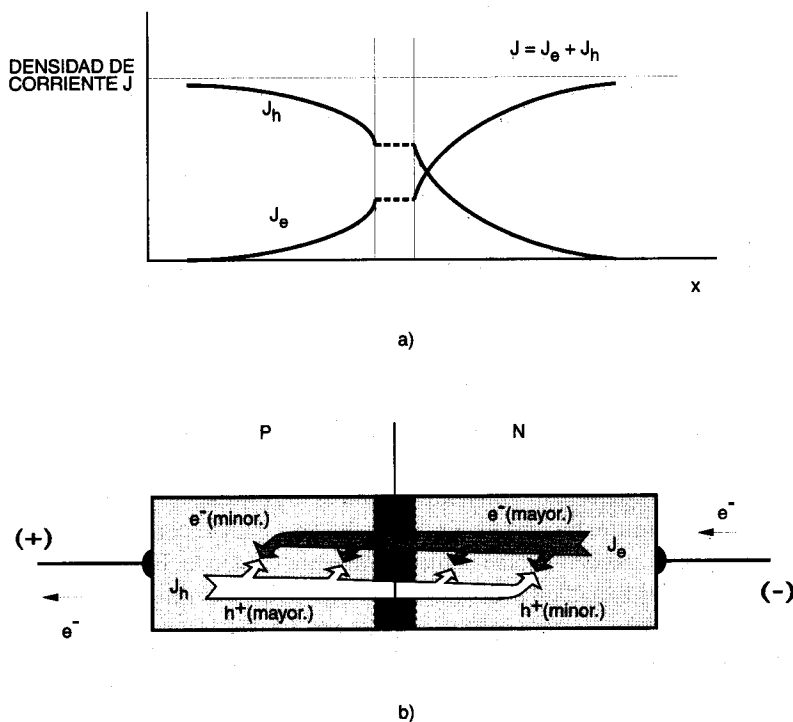


Fig.3.6. a) Variación de densidad de corriente de electrones, J_e , y de huecos, J_h , a lo largo del diodo cuando está polarizado en directo. b) Esquema del transporte de electrones y huecos en la unión p-n polarizada en directo (en el caso de polarización inversa el movimiento de los portadores se realiza en sentido contrario).

Un esquema del movimiento de los portadores y de la variación de la corriente a lo largo de la longitud del diodo (coordenada x) para el caso de polarización en directo viene dado en la fig. 3.6b. Se puede observar que la corriente en el circuito externo está transportada

únicamente por electrones. Una vez que los electrones entran por el lado n la corriente de electrones disminuye como consecuencia de los procesos de recombinación con los huecos minoritarios que proceden de la parte p. Los electrones que alcanzan la unión se difunden hacia el lado p donde a su vez se recombinan dando lugar a una nueva disminución de la corriente. El transporte de los huecos tiene un mecanismo análogo, aunque hay que tener presente que el desplazamiento de huecos se inicia en el electrodo positivo mediante un movimiento de electrones desde este electrodo hacia el circuito externo.

En el caso de polarización inversa el mecanismo de transporte de carga es similar, sin embargo en este caso la corriente se debe fundamentalmente al movimiento de portadores minoritarios a cada lado de la unión que son arrastrados por el campo (también se puede decir que se difunden) en dirección del electrodo opuesto. La concentración de portadores minoritarios es muy pequeña, y por tanto la corriente inversa será baja. Así, para voltajes inversos elevados la densidad de corriente toma el valor de J_0 , el cual está determinado por las concentraciones de equilibrio de portadores minoritarios n_{po} y p_{no} cuando no existe voltaje aplicado (ec. 3.37).

3.3.4. Régimen de ruptura

La ec. [3.36] indica que a medida que aumenta el voltaje aplicado en polarización directa la corriente a través de la unión crece según una ley cuasi-exponencial. En polarización directa, el límite superior de la corriente viene impuesto simplemente por el calor máximo que puede disipar el diodo como consecuencia del calentamiento por efecto Joule. Sin embargo, en polarización inversa la corriente a través del diodo es muy pequeña, y por ello apenas existe calentamiento. De todos modos, para voltajes en inversa suficientemente elevados existen otros procesos que limitan severamente el funcionamiento del diodo. Estos procesos, denominados *avalancha* y *túnel*, dan lugar a un aumento considerable de la corriente cuando el voltaje inverso aplicado alcanza un valor crítico. Veamos en qué consiste cada uno de estos procesos.

a) Régimen de avalancha:

Cuando se aplican al diodo voltajes inversos cada vez más elevados, los electrones minoritarios del lado p y los huecos minoritarios del lado n que entran en la región de carga espacial adquieren velocidades cada vez más elevadas como consecuencia del alto campo eléctrico presente en esta región (véase problema 3.9). Cuando el voltaje es suficientemente elevado los portadores adquieren una energía cinética capaz de ionizar por impacto a los átomos que se encuentran en reposo dentro de la región de carga espacial. En este proceso de interacción, en el que los electrones ionizan los átomos de la red, se originan nuevos electrones en la banda de conducción y huecos en la banda de valencia. Así pues, después de cada ionización existen, además del electrón primario, dos nuevos portadores, señalados como e^- y h^+ en la fig. 3.7a, que se mueven en direcciones opuestas. Tanto el electrón primario como el

electrón y el hueco generados pueden sufrir nuevas colisiones por lo que el resultado final es una avalancha de electrones en el extremo derecho de la región de carga espacial y de huecos en el extremo opuesto. La corriente a través del diodo adquiere en estas condiciones una proporción enorme con un valor mucho más elevado que el previsto en el proceso de difusión. Debido a que el campo eléctrico es más intenso justo en el centro de la unión, este proceso de generación de pares electrón-hueco tiene mayor importancia en la parte central de la región de carga espacial.

b) Proceso túnel:

Este proceso se presenta cuando el campo externo aplicado es suficientemente alto por sí solo para arrancar directamente electrones de la red de átomos en la región de carga espacial, sin mediación de colisiones (*efecto Zener*). Esto puede ocurrir cuando el estrechamiento de la banda prohibida por efecto del campo aplicado es tal que los electrones en la banda de valencia rompen su enlace con los átomos y pasan por *efecto túnel* a un estado vacío en la banda de conducción. Este fenómeno se explica por la naturaleza ondulatoria de los electrones, los cuales tienen una función de onda asociada de varias decenas de angstrom de extensión. Cuando la banda de energía prohibida se estrecha por debajo de este límite la probabilidad de ionización por efecto túnel es muy alta dando lugar a un aumento considerable de la corriente a través de la unión (fig. 3.7b).

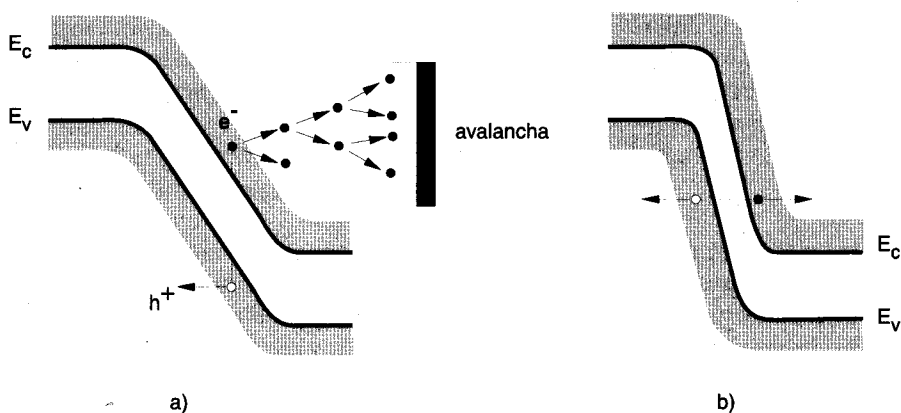


Fig.3.7. Esquema de la formación de pares electrón hueco mediante: a) ionización por impacto y posterior multiplicación en avalancha, y b) efecto túnel desde la banda de valencia a la de conducción (nótese el estrechamiento de la banda prohibida como consecuencia de la aplicación de un campo eléctrico elevado).

El efecto túnel es competitivo con el de multiplicación por avalancha de forma que solamente será aparente cuando la anchura de la región de carga espacial sea ya de partida muy estrecha. Esto ocurre por ejemplo cuando los semiconductores que componen la unión p-n tienen un dopaje muy elevado. Por contra, para un campo eléctrico dado la multiplicación por avalancha será tanto mayor cuanto más ancha sea la región de carga espacial ya que el número de colisiones ionizantes será también mayor. En cualquier caso el campo eléctrico requerido para que aparezca alguno de estos dos procesos está en el rango de 10^6 Vcm^{-1} o mayor.

A menudo se saca partido del aumento de corriente en polarización inversa como consecuencia del régimen de avalancha o túnel para ciertas aplicaciones tales como la de estabilización de fuentes de tensión utilizando los denominados *diodos Zener*, o la generación de potencia en microondas con los *diodos IMPATT*. En ambos casos se trata de diodos especiales que trabajan en el régimen de avalancha. Este tipo de aplicaciones exige un diseño especial del diodo para maximizar el aumento de corriente como consecuencia de los fenómenos de avalancha o túnel. En el capítulo V se estudia el funcionamiento de los diodos Zener así como sus posibles aplicaciones.

3.4. CURVA CARACTERISTICA INTENSIDAD-VOLTAJE DEL DIODO

La ecuación [3.36], referida como la ley del diodo en el caso ideal, está representada por la *curva característica del diodo* (véase por ejemplo la fig.3.8). Cuando se tiene en cuenta posibles efectos de recombinación en la región de carga espacial se obtiene para la corriente I , de forma aproximada:

$$I = I_0 [\exp(qV/\eta kT) - 1] \quad [3.38]$$

donde η es el llamado *factor de idealidad*, que puede variar entre 1 y 2. El valor $\eta=1$ indica que el proceso dominante de transporte a través de la unión es de difusión, caso del germanio, p.e., mientras que $\eta = 2$ se refiere a procesos de transporte en los que predomina la recombinación (para el silicio, $\eta = 2$).

La fig. 3.8a muestra la curva característica intensidad-voltaje de un diodo típico. Obsérvese que existe una variación muy abrupta en la región próxima al origen (mostrado con más detalle en el inserto de la figura), de forma que para voltajes positivos de unas décimas de voltio la corriente alcanza valores relativamente elevados, en cambio en polarización inversa la corriente es siempre muy pequeña. Hay que notar que en esta región de voltajes negativos, la corriente alcanza el valor de saturación, I_0 , para valores de unas décimas de voltio aplicadas en los extremos del diodo. Sin embargo, cuando la tensión en inversa alcanza el valor corres-

pendiente a la *tensión de ruptura*, V_R , la corriente a través del diodo aumenta abruptamente por encima del valor I_o , según se ha estudiado en el apartado 3.3.4 (fig. 3.8b).

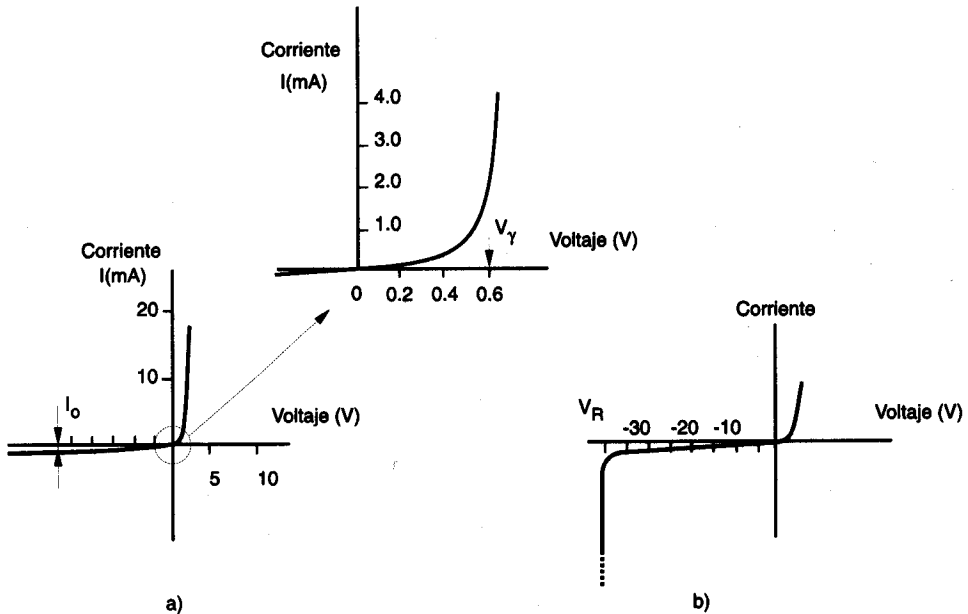


Fig. 3.8. Curva característica de un diodo típico de silicio en régimen de funcionamiento normal (a) y de ruptura (b).

Existe un voltaje de polarización en directo, V_γ , para el cual el valor de la corriente inicia un aumento muy rápido. Por debajo de V_γ la corriente es muy pequeña, inferior al 1% del valor típico para el punto de operación. Este *voltaje*, denominado de *corte o umbral*, tiene un valor de 0,6 V para el silicio y de 0,2 V para el germanio, aproximadamente. Además, la corriente en sentido directo en el silicio es mucho menor que en el germanio cuando se aplica un determinado voltaje, lo cual es debido, entre otros factores, a que la anchura de la banda prohibida es mayor en el silicio que en el germanio (véase tabla 2.2). Este hecho a su vez da lugar a que la corriente inversa de saturación I_o para el germanio ($\approx 10^{-6}$ A) sea mucho mayor que para el silicio ($\approx 10^{-9}$ A).

El aumento de la temperatura origina siempre un aumento de corriente a través del diodo. La explicación es simple: un aumento de la temperatura en un semiconductor extrínse-

co tipo n ó p da lugar a que el nivel de Fermi se acerque al nivel intrínseco, E_i . En estas condiciones la diferencia $(E_i)_n - (E_i)_p$ se hace menor por lo que la altura de la barrera, qV_o , disminuye. Al mismo tiempo, la concentración de portadores minoritarios a cada lado de la unión aumenta. En consecuencia, tanto la corriente en sentido directo como la corriente de saturación en inverso deben aumentar. Este aspecto de los diodos es crítico, ya que su cualidad más importante es el bloqueo de la corriente en sentido inverso. Se ha constatado que el *coeficiente de temperatura*¹ del Ge y del Si para la corriente de saturación es del orden de $7\% / ^\circ\text{C}$, lo cual indica que I_o se duplica por cada aumento de la temperatura en una década.

Debido a la variación quasi-exponencial de la intensidad con el voltaje, la *resistencia estática* del diodo, esto es el cociente V/I , siendo I la corriente total, tiene un valor dependiente del voltaje aplicado. Por ello es más frecuente usar en las especificaciones el valor del voltaje en directo, V_D , necesario para alcanzar una determinada corriente, I_D , 10 mA por ejemplo. Asimismo, se suele dar dentro de las especificaciones la corriente de saturación en polarización inversa, I_o .

Cuando se trabaja con pequeñas señales de voltaje superpuestas a una tensión continua, el voltaje varía en una pequeña cantidad, ΔV , alrededor de un valor constante, V . Se suele dar entonces como parámetro característico del diodo la *resistencia incremental o dinámica*,

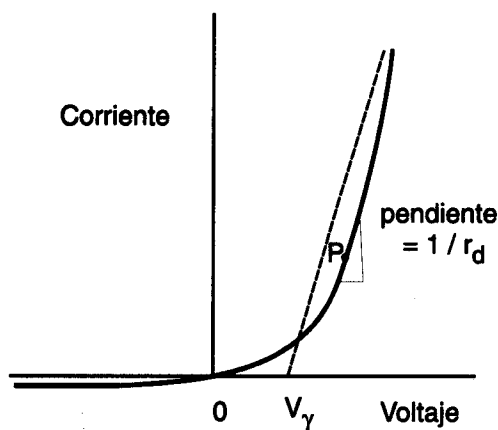


Fig.3.9. Cálculo de la resistencia dinámica en el punto P de funcionamiento del diodo. Se ha representado mediante trazos la aproximación de la curva característica del diodo por dos tramos lineales.

¹ Nota: El coeficiente de temperatura para la corriente se define como el porcentaje de variación de la corriente por cada grado de temperatura aplicado a la unión.

definida como $r_d = (dV/dI)_V$, la cual depende obviamente del voltaje V en el punto de operación. El valor de r_d coincide con el inverso de la pendiente de la curva característica del diodo en el punto de operación (fig. 3.9). De la ec. [3.38] se encuentra:

$$r_d = \frac{dV}{dI} = \frac{1}{dI/dV} = \frac{\eta kT}{q I_0 \exp(qV/\eta kT)} = \frac{\eta kT}{q (I_0 + I)} \quad [3.39]$$

Para voltajes en directo superiores a unas décimas de voltio, $I \gg I_0$, por tanto:

$$r_d \approx \frac{\eta kT}{q I} = \frac{\eta V_T}{I} \quad [3.40]$$

siendo V_T el voltaje equivalente de la temperatura definido en el apartado 2.5. En polarización en directo, el valor de r_d es siempre pequeño, en la región de unos pocos ohmios. En cambio para voltajes inversos, la resistencia r_d es en general muy elevada. A menudo, desde un punto de vista práctico, se hace una aproximación lineal por tramos de la característica $I - V$ del diodo sustituyéndolo por una resistencia infinita para voltajes inferiores al umbral y por una resistencia constante igual al valor de r_d (para el voltaje de trabajo) en la región de voltajes superiores al umbral (línea a trazos en la fig. 3.9). En la región de ruptura el diodo se sustituye también a veces por una resistencia de valor cero (no representada en la figura).

3.5. CAPACIDAD ASOCIADA A LA UNION

Una vez estudiado el comportamiento del diodo como elemento resistivo interesa estudiar también sus aspectos capacitivos. Efectivamente, según se ha visto en el diodo existe una cierta carga estática, Q , igual y de signo opuesto, distribuida a cada lado de la unión en la región de carga espacial. Al mismo tiempo, cuando se aplica un voltaje a los extremos del diodo, existe también un exceso (o defecto) de carga debida a los portadores minoritarios (se trata de carga libre en este caso) en las zonas de la región neutra que están próximas a la región de carga espacial. Tanto la carga estática, Q , como el exceso (o defecto) de carga libre dependen del voltaje aplicado y contribuyen conjuntamente a los fenómenos capacitivos asociados a la unión.

3.5.1 Carga acumulada en la región de carga espacial

En la región de carga espacial existe una carga fija cuyo valor total por unidad de superficie, Q_n' y Q_p' , en los lados n y p , respectivamente, viene dado por:

$$Q_n' = Q_n/S = q N_d x_n \quad [3.41]$$

y

$$Q_p' = Q_p / S = -q N_a x_p \quad [3.42]$$

siendo S la superficie transversal de la unión. Cuando no existe voltaje aplicado, x_n y x_p vienen dados por las ecs. [3.12] y [3.13], respectivamente. Para obtener los correspondientes valores de x_n y x_p cuando se aplica un voltaje externo es preciso generalizar dichas expresiones sustituyendo el valor de V_o por $(V_o - V)$, como se hizo en el apartado 3.2. Para el lado n de la región de carga espacial tendremos por ejemplo:

$$x_n = \left[\frac{2\epsilon (V_o - V)}{q N_d} \left(\frac{N_a}{N_d + N_a} \right) \right]^{1/2} \quad [3.43]$$

con un valor similar para el lado p .

De las ecuaciones anteriores se desprende que la carga acumulada en la región de carga espacial no varía de forma lineal con el voltaje externo. Por ello es preciso definir la capacidad asociada a esta región, C_s , como la variación de la carga en uno de los lados (en el lado n por ejemplo) en relación a cambios pequeños en el voltaje, es decir : $C_s = dQ_n/dV$. Por unidad de superficie tendremos:

$$C_s' = \frac{dQ_n'}{dV} \quad [3.44]$$

Dado que Q_n' varía con el espesor x_n y éste, a su vez, varía con el voltaje, podremos poner:

$$C_s' = \frac{dQ_n'}{dV} = \frac{dQ_n'}{dx_n} \left| \frac{dx_n}{dV} \right| \quad [3.45]$$

En esta expresión se ha tomado el valor absoluto del último factor para evitar un resultado negativo en la capacidad, ya que el valor de dx_n/dV es negativo. Calculando las derivadas de las ecs. [3.41] y [3.43] y utilizando la ec. [3.45], resulta finalmente para la capacidad por unidad de área:

$$C_s' = \frac{\epsilon}{x} \quad [3.46]$$

Esta relación indica que el diodo se comporta como un condensador ideal de láminas plano-paralelas con un espesor, x , igual al espesor total de la región de carga espacial y con la carga localizada en los extremos de dicha región.

En una unión abrupta de tipo p⁺-n (es decir con $N_a \gg N_d$), el espesor total de la región de carga espacial viene determinado fundamentalmente por el espesor correspondiente al lado menos dopado (el lado n en este caso). Así, de las expresiones [3.16] y [3.43] se obtiene haciendo $N_a \gg N_d$:

$$x \approx x_n = \left[\frac{2\epsilon (V_o - V)}{q N_d} \right]^{1/2} \quad [3.47]$$

Para uniones de este tipo, se puede obtener una expresión simple que relaciona la variación de la capacidad del diodo con el voltaje aplicado. Combinando las expresiones [3.46] y [3.47] resulta:

$$\frac{1}{C_s^2} \approx \frac{2(V_o - V)}{\epsilon q N_d} \quad [3.48]$$

Por tanto, si se mide la variación de la capacidad de la unión con el voltaje se puede obtener, mediante la representación gráfica de $1/C_s^2$ frente a V, los valores de N_d y V_o a partir de la pendiente de la recta y de la abscisa en el origen. Este tipo de medidas son muy utilizadas para la caracterización de los semiconductores.

3.5.2. Carga acumulada en las regiones neutras

Cuando la unión se polariza en directo existe otra contribución a la capacidad que puede modificar sensiblemente el valor de la capacidad asociada a la carga espacial, dada por la ecuación [3.46]. Se trata en este caso de la *capacidad de difusión* que es debida al exceso de portadores minoritarios, de tipo p ó n, acumulados en la región neutra a ambos lados de la región de carga espacial y que proceden de la difusión desde el lado opuesto (fig. 3.10). La concentración de este exceso de portadores minoritarios ha sido calculada en las expresiones [3.30] y [3.31], las cuales muestran una variación exponencial de los portadores a medida que nos alejamos de la unión, según refleja la fig. 3.5. El cálculo de la carga total acumulada se puede hacer, por tanto, mediante integración simple de las ecs. [3.30] y [3.31], extendiendo los límites de la integral desde el borde de la región neutra hasta el extremo correspondiente del diodo. El resultado final para la densidad de carga asociada a los huecos acumulados en el lado n es:

$$Q_h' = q L_n p_{no} [\exp(qV/kT) - 1] \quad [3.49]$$

y una expresión similar para la carga de los electrones en el lado p. Comparando esta expresión con el valor de la densidad de corriente de huecos en el borde de la región neutra,

ec. [3.35], se observa que:

$$Q_h' = \frac{L_h^2}{D_h} J_h(0) = \tau_h J_h(0) \quad [3.50]$$

Es decir, la carga en exceso por unidad de superficie acumulada en el lado n viene dada por el producto de la densidad de corriente en el borde de la región neutra por el tiempo de vida media de los portadores, τ_h , esto es, el tiempo medio que tardan los portadores minoritarios en recombinarse una vez que han atravesado la unión. Este resultado es explicable si se tiene en cuenta que la carga acumulada será mayor cuanto mayor sea el tiempo de vida media. Un resultado similar se obtiene para los electrones acumulados en el lado p. Se puede concluir por tanto que cuando un diodo se polariza en sentido directo la corriente que pasa por el diodo suministra portadores minoritarios a cada lado de la unión al mismo ritmo que éstos desaparecen, manteniendo constante la carga total acumulada en la zona neutra.

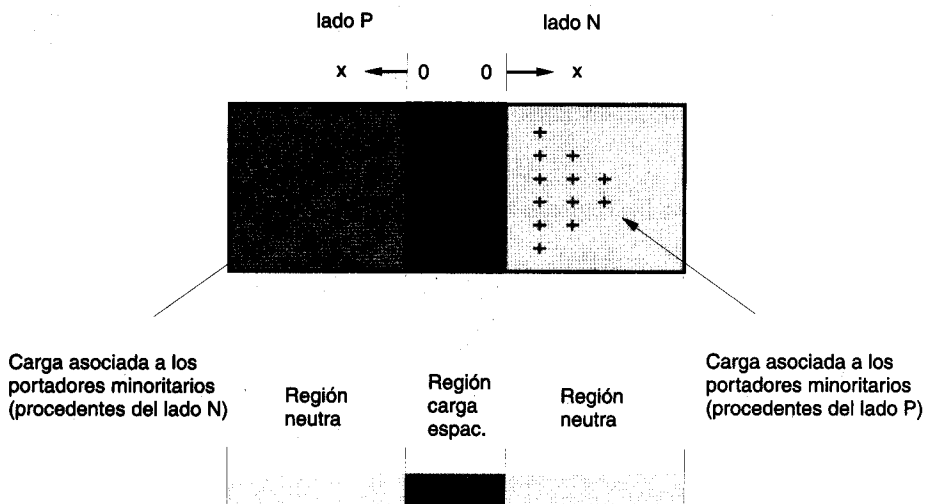


Fig.3.10. Distribución de carga debida al exceso de portadores minoritarios acumulados en la regiones neutras de una unión p-n polarizada en directo.

La capacidad de difusión, C_d , se puede obtener mediante la expresión $C_d = dQ/dV$, utilizando para Q el valor de la carga en exceso acumulada en cada uno de los lados. Para el lado n, por ejemplo, $Q = Q_h'S$, siendo S la superficie de la unión. Imponiendo la condición

adicional de que la carga almacenada en cada lado debe ser la misma, $\tau_h J_h(0) = \tau_e J_e(0)$, y utilizando la ecuación [3.36], resulta tras un sencillo cálculo (véase problema 3.12):

$$C_d = \frac{dQ_h S}{dJ} \frac{dJ}{dV} = \tau S \frac{dJ}{dV} = \tau \frac{dI}{dV} = \frac{\tau}{r_d} \quad [3.51]$$

siendo r_d la resistencia dinámica del diodo definida mediante la expresión [3.39] y τ una constante con unidades de tiempo definida por: $1/\tau = (1/\tau_e + 1/\tau_h)$. La relación [3.51] es válida también para el caso de polarización en sentido inverso, aunque entonces r_d es tan elevada que el valor de C_d puede despreciarse. En este caso la mayor contribución a la capacidad procede de la capacidad asociada a la carga espacial.

Los aspectos capacitivos de la unión p-n indican que el diodo tiene un comportamiento complejo cuando se aplican señales alternas en sus extremos. Por esta razón, y para simplificar el análisis de su comportamiento en circuitos electrónicos, muy a menudo se sustituye el diodo por el *circuito equivalente*, formado por la asociación de resistencias o condensadores discretos de valor conocido. Cuando el diodo está funcionando con una pequeña señal alterna superpuesta a un voltaje V de polarización en continua, el diodo puede ser sustituido por una

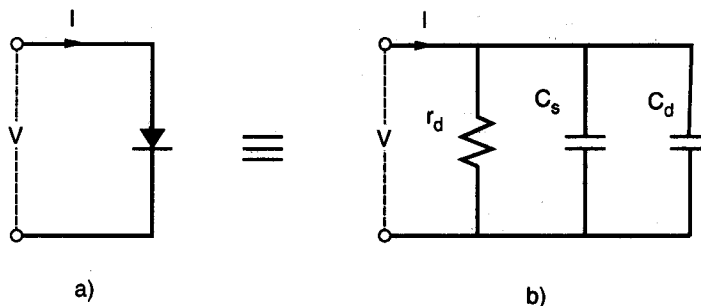


Fig.3.11. a) Símbolo empleado en circuitos para representar un diodo. b) Circuito equivalente de pequeña señal de un diodo de unión.

resistencia y dos condensadores en paralelo, tal como se indica en la fig. 3.11. La resistencia r_d corresponde a la resistencia dinámica de la unión, dada por la ec. [3.39], mientras que los condensadores C_s y C_d están asociados a la capacidad de la región de carga espacial y a la

capacidad de difusión, ecs. [3.46] y [3.51], respectivamente, todos ellos con un valor variable dependiente del voltaje aplicado.

Los aspectos capacitivos de la unión p-n son aprovechados frecuentemente en la tecnología de circuitos integrados para preparar condensadores. En este tipo de aplicaciones, el diodo se polariza con un voltaje en inverso con objeto de que la corriente a través de él sea muy baja. De este modo, el comportamiento del diodo se aproxima más al de un condensador ideal. Aunque la capacidad que se puede conseguir por este procedimiento no es muy elevada, sin embargo existen grandes dificultades de integrar en pequeñas áreas otros tipos de condensadores más convencionales.

3.5.3. Tiempo de conmutación del diodo (*)

La capacidad de difusión a su vez determina el tiempo de conmutación del diodo. El tiempo de conmutación tiene mucha importancia en circuitos digitales, ya que en este tipo de aplicaciones el diodo funciona generalmente en dos estados de polarización a un voltaje fijo en cada uno de ellos: polarización directa (estado "on") y polarización inversa (estado "off"). El tiempo de conmutación es el tiempo que tarda el diodo en pasar de un estado a otro. Obviamente, interesa que este tiempo sea lo más corto posible con objeto de aumentar la velocidad de operación del circuito.

En estos cambios de estado los portadores mayoritarios a cada lado de la unión responden de manera prácticamente instantánea, ya que su distribución espacial no varía de forma sensible con la tensión aplicada. Por esta razón se puede considerar que la carga espacial responde también rápidamente a los cambios de voltaje. No ocurre, sin embargo, lo mismo con los portadores minoritarios almacenados en las zonas neutras, adyacentes a la carga espacial, ya que en este caso la distribución de los portadores cambia notablemente (véase p.e. figs. 3.5a y 3.5b, parte inferior). Para visualizar este efecto, consideremos un diodo de unión abrupta del tipo p⁺-n polarizado en directo. Si la corriente que circula es J_{dir} la densidad de carga Q' acumulada en el lado n vendrá dada de forma aproximada por la expresión [3.49]. Si en el instante $t = 0$ cambiamos la polaridad aplicada al diodo, inmediatamente aparece un campo eléctrico en la unión que se añade al propio campo eléctrico de la carga espacial. El exceso de huecos minoritarios que estaban acumulados en el lado n se difunden hacia el lado p ayudados por el campo eléctrico en la unión. Instantáneamente, la densidad de corriente en sentido inverso, J_{inv} , será grande (fig. 3.12), pero a medida que desaparece la carga acumulada, Q' , la corriente en sentido inverso será cada vez menor, de forma que el diodo se acerca al estado estacionario, en el que la corriente tiene un valor muy pequeño, cada vez más lentamente. El tiempo de conmutación se define como el tiempo necesario para alcanzar un determinado valor, próximo al equilibrio. Este tiempo de conmutación depende de muchos factores, entre ellos de la densidad de carga total acumulada, Q' , durante el período de funcionamiento en directo. Dicha carga es función exponencial del voltaje (ec. 3.49). Por la misma razón, el tiempo de conmutación también dependerá del tiempo de vida media de los portadores mino-

ritarios, τ . En aquellas aplicaciones donde se pretende un tiempo de conmutación corto, interesa diseñar el diodo de forma que τ sea lo más bajo posible, por ejemplo introduciendo en el diodo impurezas que actúan como centros de recombinación (sec.2.6.1). El tiempo de conmutación en estos casos puede ser de unos 10^{-10} s. Sin embargo los valores típicos en un diodo normal son mucho mayores, pudiendo alcanzar hasta 10^{-5} s.

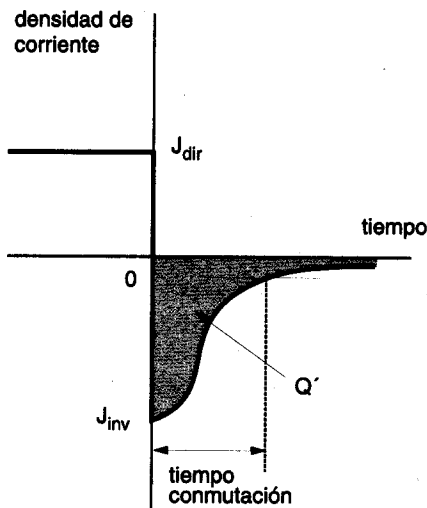


Fig.3.12. Diagrama cualitativo de la variación de la corriente al cambiar el estado de polarización de un diodo desde el sentido directo al inverso.

CUESTIONES Y PROBLEMAS

- 3.1 En una unión abrupta de Si, con una concentración de impurezas aceptoras de 1 átomo por cada 10^8 átomos de Si y de impurezas donadoras, $N_d = N_a \times 10^3$, calcular: 1º) el potencial de contacto, V_o , 2º) la concentración de minoritarios a cada lado de la unión.
- 3.2 La concentración de impurezas en un diodo de Si es $N_a = 5 \times 10^{15} \text{ cm}^{-3}$ y $N_d = 10^{16} \text{ cm}^{-3}$. Suponiendo que el tiempo de vida media de los portadores es $\tau_e = \tau_h = 1 \text{ } \mu\text{s}$, calcular: 1º) la longitud de difusión de los portadores, 2º) la variación de la concentración de

minoritarios con la distancia a cada lado de la unión para un voltaje aplicado de 0,5 V, a la temperatura ambiente. ¿Hasta qué punto es válida la condición de baja inyección?, 3º) calcular la densidad de corriente de saturación, J_o ($D_e = 33.8 \times 10^{-4} \text{ m}^2 \text{ s}^{-1}$ y $D_h = 13.0 \times 10^{-4} \text{ m}^2 \text{ s}^{-1}$).

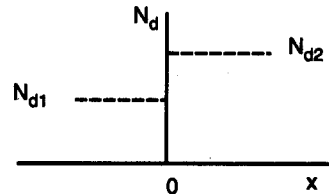
- 3.3** Con los datos del problema anterior, hacer una representación gráfica de la característica I - V a 100 y 300 K suponiendo que la superficie del diodo es $4 \times 10^{-2} \text{ cm}^2$ (utilícese para N_e y N_v los resultados del problema 2.3).
- 3.4** Un diodo de Si está dopado con $N_a = 5 \times 10^{15} \text{ cm}^{-3}$ y $N_d = 10^{15} \text{ cm}^{-3}$. Suponiendo que $\tau_h = 0.4 \mu\text{s}$ y $\tau_e = 0.1 \mu\text{s}$, calcular: 1º) la corriente de saturación debida a los electrones y a los huecos, 2º) las concentraciones en los bordes de la región de agotamiento y a una distancia del borde igual a la longitud de difusión, al aplicar una tensión igual a $V_o / 2$ (utilícese para D_e y D_h , los datos del problema 3.2).
- 3.5** Demostrar que las contribuciones J_e y J_h (corriente debida a electrones y huecos, respectivamente) a la corriente de saturación, $J_o = J_e + J_h$, cumplen la relación: $J_e / J_h = \sigma_e L_h / \sigma_h L_e$, siendo σ_e y σ_h las conductividades de las regiones n y p, respectivamente.
- 3.6** El diodo del problema 3.3 se polariza en directo a 300 K con una fuente de alimentación de 1.0 V a través de una resistencia en serie de 100 ohm. Calcular el punto de operación del diodo y la resistencia dinámica en dicho punto. Resolver el mismo problema con el diodo polarizado en inverso.
- 3.7** Calcular la característica I-V de un diodo que tiene una corriente de saturación de 10 mA, con una resistencia de las regiones n y p de 25 ohm. Compárese el resultado con la característica I-V del diodo ideal.
- 3.8** Calcular el coeficiente de temperatura de la corriente a través de un diodo suponiendo que el coeficiente de difusión varía con la temperatura según una ley inversa. Aplíquese al caso de un diodo de Si polarizado en directo a un voltaje de 0.6 V.
- 3.9** Demostrar que el campo eléctrico $E(0)$ en el centro de una unión p-n varía con el voltaje externo aplicado, V, según la ecuación:

$$E(0) = - \left[\frac{2q(V_o - V)}{\epsilon} \left(\frac{N_a N_d}{N_a + N_d} \right) \right]^{1/2}$$

- 3.10** Una unión p-n de Si está dopada con $N_a = 10^{16} \text{ cm}^{-3}$ y $N_d = 10^{15} \text{ cm}^{-3}$ átomos de impurezas. Sabiendo que la constante dieléctrica del silicio es 11.8. Calcular: 1º) el potencial de contacto V_o , 2º) la anchura total de la región de carga espacial, 3º) las anchuras

correspondientes a los lados n y p, y 4°) el campo eléctrico en el centro de la unión ($x=0$).

- 3.11** Calcular la variación del campo eléctrico y el potencial en la región de carga espacial de una unión p-n de dopaje gradual. Demostrar que la capacidad de la unión varía con el voltaje aplicado según una ley del tipo $1/C \propto V^{1/3}$ (la unión de dopaje gradual, ó *union gradual* es aquella en la que el dopaje a cada lado de la unión, $N(x)$, varía con la distancia al centro siguiendo una ley del tipo $N(x) = kx$).
- 3.12** Demostrar que el exceso de carga por unidad de superficie debido a los portadores minoritarios acumulados en las regiones neutras de una unión p-n polarizada a un cierto voltaje viene dado por $Q_h' = \tau_h J_h(0)$ para los huecos en el lado n. Deducir también el valor de la capacidad de difusión $C_d = \tau/r_d$ (ec. 3.51).
- 3.13** Una unión de silicio abrupta está dopada a cada lado de la unión solamente con impurezas donadoras (dopaje isotópico) con concentraciones N_{d1} y N_{d2} (ver figura). Si $N_{d1} < N_{d2}$: 1º) Calcular el potencial de contacto, V_o , y hacer un esquema de la curvatura de las bandas de energía, 2º) Hacer un diagrama cualitativo de la variación de la densidad de carga, del campo eléctrico, y del potencial a lo largo de la unión.



CAPITULO V

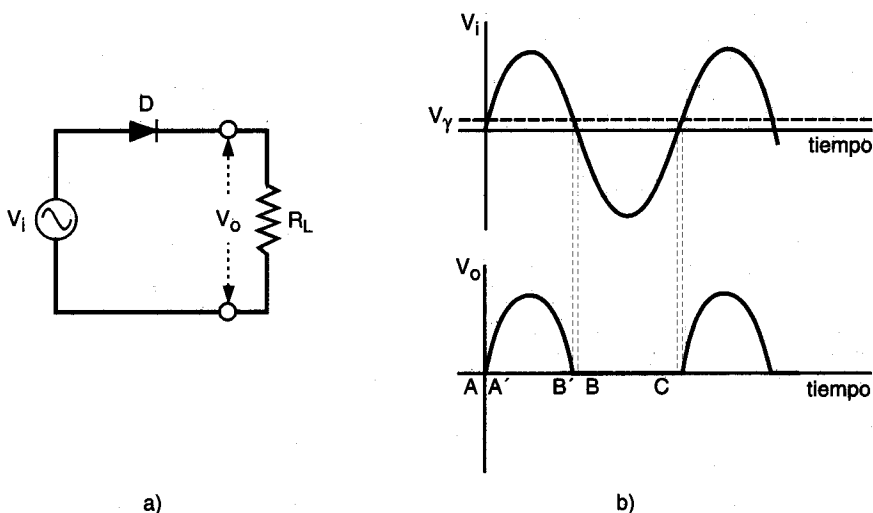
APLICACIONES DE LOS DIODOS SEMICONDUCTORES

Desde el descubrimiento del efecto rectificador de la unión p-n en la década de los 40, se ha desarrollado una gran variedad de dispositivos cuyo funcionamiento está basado fundamentalmente en las propiedades de la unión p-n. Entre ellos, cabe destacar sobre todo los dispositivos opto-electrónicos, tales como los fotodetectores y los diodos emisores de luz, así como los modernos diodos láser. Entre las muchas aplicaciones de estos dispositivos hay que citar por ejemplo el campo de las comunicaciones por fibra óptica, mediante las cuales es posible transmitir mucha mayor información que en los sistemas de comunicación por cable convencional. En este capítulo se trata de dar una visión general de los fundamentos físicos de los diferentes dispositivos basados en la unión p-n, sin entrar en detalle en los circuitos electrónicos, más o menos complejos, desarrollados para sacar mejor partido de ellos.

5.1. EL DIODO COMO ELEMENTO RECTIFICADOR

Una de las aplicaciones más inmediatas de los diodos es la rectificación de tensiones alternas. Según se ha visto en el capítulo tercero, los diodos semiconductores presentan una elevada resistencia cuando están polarizados en inverso, mientras que su resistencia es prácticamente cero cuando se les polariza en directo con un voltaje superior al umbral, V_y (véase apartado 3.4).

En la fig. 5.1a, se presenta un circuito simple de rectificación de una tensión alterna, V_i , en el que se incluye un diodo en serie con una resistencia, R_L . Esta resistencia representa un elemento de consumo sobre el que se pretende aplicar tensiones únicamente positivas en uno de sus extremos. En el semiperíodo positivo de la señal del generador, V_i , el diodo presenta



una resistencia muy pequeña, por lo que prácticamente toda la tensión cae en la resistencia externa de carga R_L . La onda de salida, V_o , es entonces prácticamente igual a la señal de entrada (tramo AB en la fig. 5.1b). En contraste, durante el semiperíodo negativo de la señal V_i el diodo ofrece una resistencia muy elevada, de forma que para efectos prácticos puede considerarse infinita. En estas circunstancias, toda la tensión V_i cae en el diodo y en consecuencia la tensión V_o que cae en R_L es próxima a cero (tramo BC de la fig. 5.1b). Así pues, mediante el circuito simple de la fig. 5.1a sólo los ciclos positivos de la señal actúan sobre la carga, y por esta razón se le denomina *rectificador de media onda*. Debido a que para tensiones positivas por debajo de la tensión umbral la resistencia del diodo es ya muy elevada, una pequeña parte del ciclo positivo (tramos AA' y BB' en la fig. 5.1b) aparece también anulada en la señal de salida. Es evidente además, que el conjunto diodo-resistencia de carga actúa como un *divisor de tensión* de forma que el efecto rectificador está limitado a resistencias de carga mucho menores que la resistencia del diodo en polarización inversa, y a la vez mucho mayores que la resistencia en directo.

Una rectificación mucho más eficiente se consigue utilizando circuitos más complejos, como el *rectificador de onda completa* presentado en la fig. 5.2a. En este caso es preciso

pasar primero la tensión a rectificar, V_i , a través de un transformador. El transformador tiene en el arrollamiento del secundario una toma intermedia conectada en el punto medio de la bobina. De este modo, la diferencia de potencial en cualquier instante entre este terminal y uno cualquiera de los extremos del transformador debe ser la mitad de la tensión entre los extremos. La resistencia de carga se conecta entonces entre el terminal intermedio, punto A, conectado a tierra (cero), y el punto B común a los dos diodos, los cuales a su vez están unidos por el otro extremo a cada uno de los terminales de salida del transformador.

Debido a esta toma intermedia conectada a tierra, las tensiones entre los terminales 1-A y 2-A oscilan en oposición de fase. Así, por ejemplo, consideremos el ciclo positivo de la onda de entrada en el cual la tensión en los extremos del transformador pasa por un valor máximo, V_m . El voltaje máximo entre los puntos 1 y A (tierra) será $+V_m/2$, mientras que entre 2 y A el voltaje máximo será de $-V_m/2$. De este modo, la diferencia de potencial $V_{12} = V_{1A} - V_{2A}$ coincide con el valor V_m . Por tanto, durante el semiciclo positivo el diodo D_1 queda polarizado en directo y transmite la señal positiva entre los terminales A y B, mientras que el diodo D_2 permanece polarizado en inverso, "bloqueando" la señal negativa. Como consecuencia de esto, entre los puntos A y B donde está conectada la carga R_L , aparece una onda o señal de amplitud $+V_m/2$. Durante el período negativo de la onda de entrada ocurre una situación similar, aunque ahora tendremos una señal alterna de valor máximo $+V_m/2$ entre los puntos 2 y A (tierra) y $-V_m/2$ entre los puntos 1 y A. Tenemos pues que en este semiperíodo el diodo D_1 "bloquea" la señal, mientras que el diodo D_2 transmite la señal positiva hasta el

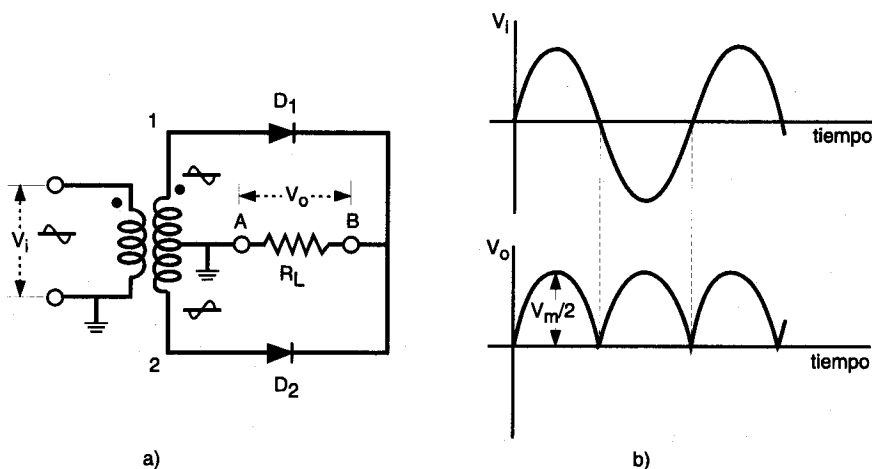


Fig. 5.2. a) Circuito rectificador de onda completa. b) Esquema de la variación de la señal de salida, V_o , para una onda sinusoidal de entrada, V_i .

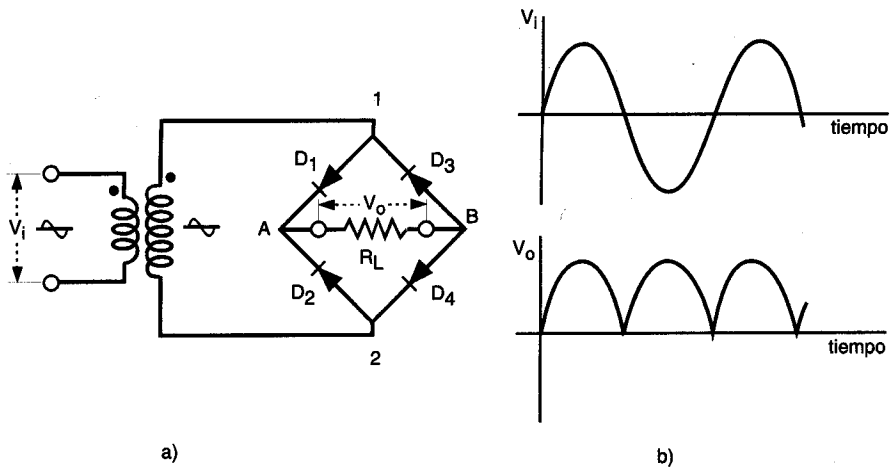


Fig. 5.3. a) Circuito rectificador de onda completa formado por un puente de cuatro diodos. b) Variación de las ondas de entrada, V_i , y de salida, V_o .

punto B. La onda de salida entre los puntos A y B tiene de nuevo un valor positivo. En resumen, durante un ciclo completo de la onda de entrada, la variación del voltaje de salida es siempre positiva, con ciclos alternantes de frecuencia doble a la de entrada según se indica en la gráfica de la fig. 5.2b.

Otro circuito rectificador de onda completa, formado por un puente de cuatro diodos, viene esquematizado en la fig. 5.3a. En este caso la carga R_L se conecta entre los extremos A y B del puente, mientras que la onda a rectificar se aplica entre los otros dos extremos, 1 y 2. Durante el ciclo positivo, el terminal 1 es positivo y el 2 es negativo por lo que la corriente se transmite a la carga R_L mediante los diodos D_1 y D_4 , que están polarizados en directo. Por contra, durante el ciclo negativo, el terminal 2 es positivo y el 1 negativo. La corriente se transmite ahora hacia la carga a través de los diodos D_2 y D_3 , que son los que están polarizados en directo. Nótese que en ambos ciclos la tensión entre los puntos A y B es siempre positiva e igual a la tensión de salida V_o (fig. 5.3b). Este circuito tiene de ventaja, sobre el anterior, de que para una señal alterna dada a la salida del transformador las tensiones máximas en inverso a que están sometidos cada uno de los diodos son más bajas. Además, este circuito no requiere el uso de un transformador, aunque suele ser frecuente su utilización en aplicaciones de baja tensión.

5.2. CIRCUITOS LIMITADORES (*)

Basadas en el efecto de rectificación, son muchas las funciones que pueden ejecutar los diodos semiconductores, sobre todo cuando van asociados con otros componentes del circuito, tales como condensadores, baterías, etc. Entre estas funciones cabe citar las de detección de valor de pico en una señal variable, multiplicación de voltajes alternos, limitación de tensiones variables, estabilización de fuentes de alimentación, etc. Por su importancia e interés haremos una descripción solamente de estas dos últimas aplicaciones.

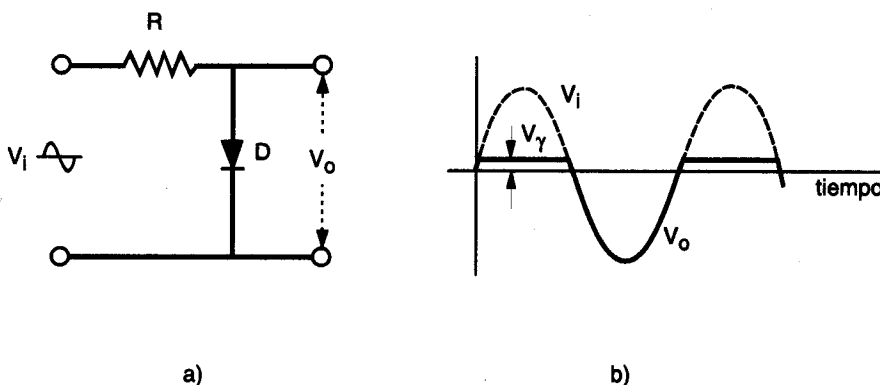


Fig. 5.4. a) Circuito limitador de señales de voltaje positivas. b) Variación de la onda de salida, V_o , para una onda sinusoidal de entrada, V_i .

A menudo interesa, por ejemplo, eliminar la parte de una tensión variable con forma arbitraria que esté por encima de un valor de referencia dado, V_r . Esto ocurre por ejemplo cuando se pretende proteger un equipo o un dispositivo de sobretensiones que eventualmente se puedan presentar sobre él.

Un ejemplo característico de limitación de voltajes lo constituyen los circuitos rectificadores estudiados en el apartado anterior ya que de hecho sólo permiten el paso de la parte positiva de un voltaje alterno. En la fig. 5.4a se da, asimismo, un circuito de similares características al circuito rectificador que sólo permite el paso de la parte negativa de un voltaje alterno, según se indica en la fig. 5.4b. Así, en el semiciclo positivo, el diodo polariza-

do en directo tiene una resistencia muy pequeña, por lo que la caída de potencial a través de él, igual a V_o , será prácticamente nula. En el semiciclo negativo el diodo se encuentra polarizado en inverso y su resistencia es elevada. En consecuencia, toda la tensión de entrada cae prácticamente a través del diodo, con lo que $V_o \approx V_i$. La fig. 5.4b ilustra el comportamiento de la onda de salida para un ciclo completo de la tensión de entrada. Obsérvese que durante el semiciclo positivo el voltaje de salida no es completamente nulo ya que el diodo no presenta una resistencia nula hasta que la tensión aplicada sobre él sobrepasa la pequeña tensión umbral, V_r .

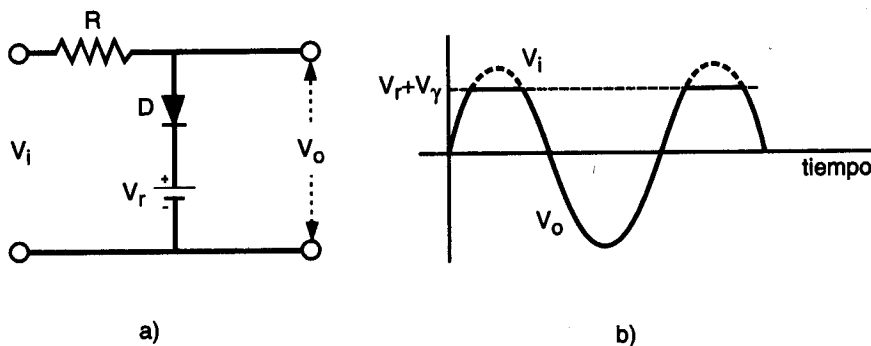


Fig. 5.5. a) Circuito limitador, de corte superior. b) Variación de la onda de salida, V_o , para una onda sinusoidal de entrada, V_i .

En las figs. 5.5a y 5.6a se dan otros ejemplos de circuitos limitadores basados en el mismo principio de funcionamiento que el circuito anterior. En ambos casos el diodo está conectado en serie con una batería cuyo voltaje V_r sirve como referencia. En el primer circuito, fig. 5.5a, el diodo está polarizado en inverso durante el período de tiempo en el cual la tensión de entrada V_i es menor que V_r . Por tanto, la resistencia del diodo es entonces muy grande y la onda de salida V_o es igual a la de entrada. En el instante en que $V_i > V_r$ (en realidad es $V_i > V_r + V_\gamma$) el diodo queda polarizado en directo con resistencia prácticamente nula. La tensión de salida V_o se iguala entonces al voltaje de referencia (o más exactamente $V_o = V_r + V_\gamma$). La forma de la onda de salida viene esquematizada en la fig. 5.5b, donde se puede observar el efecto de limitación del voltaje de entrada a un valor inferior a V_r el cual puede ser fijado de antemano.

El circuito de la fig. 5.6a funciona de la misma manera con la única diferencia de que, al estar el diodo invertido respecto de la situación anterior, la parte de la onda que queda eliminada es aquella para la cual el voltaje es inferior a V_r (fig. 5.6b). La acción combinada de

los circuitos limitadores de la parte superior y de la parte inferior de un voltaje variable se puede llevar a la práctica en un circuito único, según se esquematiza en las figs. 5.7a y 5.8a.

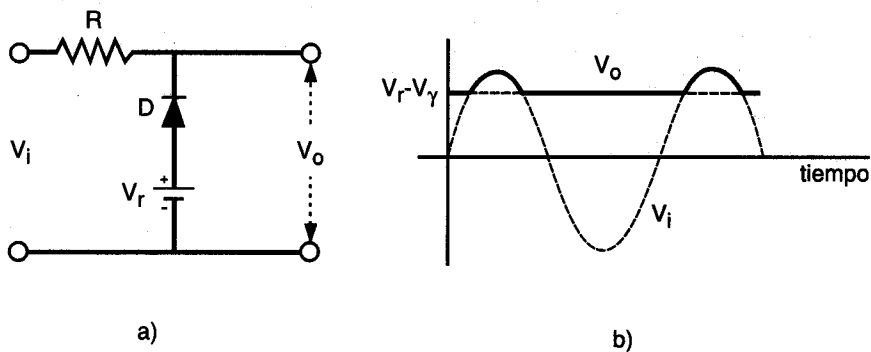


Fig. 5.6. a) Circuito limitador de corte inferior. b) Variación de la onda de salida, V_o , para una onda sinusoidal de entrada, V_i .

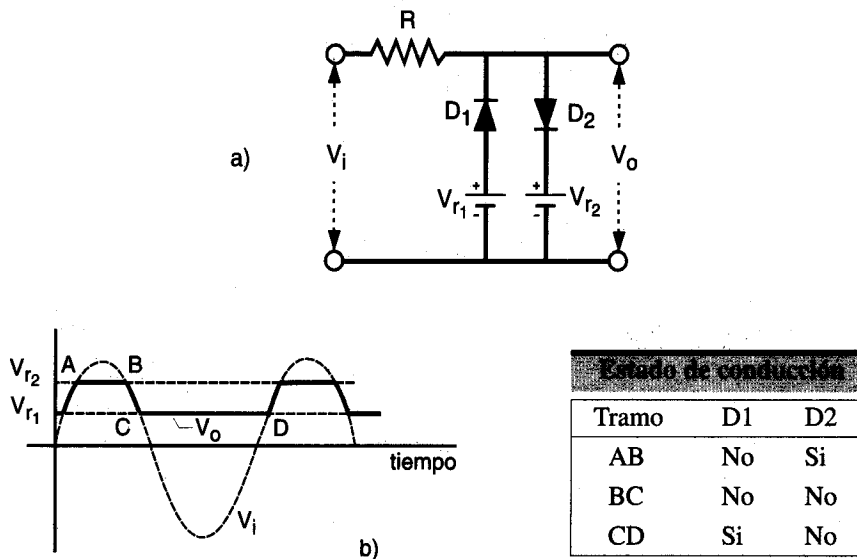


Fig. 5.7. a) Circuito limitador, de corte superior en ambos sentidos. b) Variación de la onda de salida, V_o , para una onda sinusoidal de entrada, V_i (se incluye una tabla que refleja el estado de conducción de cada diodo en cada tramo de la curva de entrada).

En ambos casos se obtiene, para una onda sinusoidal de entrada, una onda de salida recortada en los valores de cresta (figs. 5.7b y 5.8b). Eligiendo adecuadamente los valores de las tensiones de referencia se puede obtener ondas prácticamente cuadradas (fig. 5.8b). El circuito funciona entonces como un convertidor de onda sinusoidal a onda cuadrada.

5.3. ESTABILIZADORES DE TENSION: DIODOS ZENER

Las fuentes de alimentación de voltaje suelen tener un rango limitado de utilización en el régimen de voltaje constante. Según se indica en el Apéndice A3 esta limitación está impuesta por la propia resistencia interna de la fuente. Efectivamente, consideremos la fuente de voltaje de la fig. 5.9a formada por un generador, V_i , con una resistencia interna, R_i . El voltaje

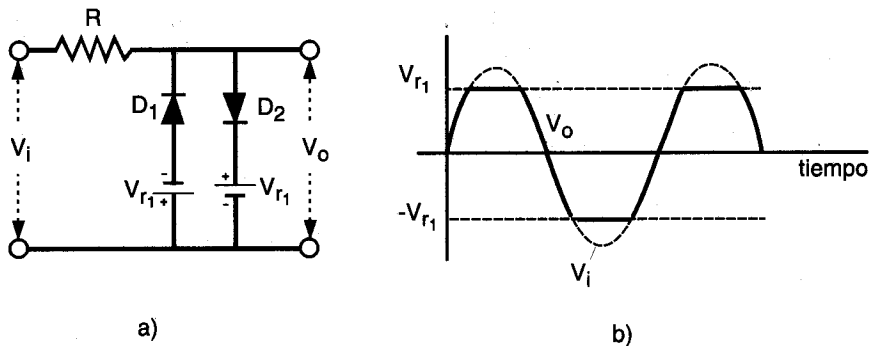


Fig. 5.8. a) Circuito limitador, de corte simétrico. b) Variación de la onda de salida, V_o , para una onda sinusoidal de entrada, V_i .

de salida de la fuente, V_o , cuando se conecta en los terminales de salida una resistencia de carga, R_L , viene dado por la ecuación:

$$V_o = V_i - IR_i \quad [5.1]$$

donde I es la corriente a través de la resistencia de carga, con un valor determinado por $I = V_i / (R_L + R_i)$. La ec. [5.1] indica que la variación de V_o en función de I corresponde a una recta con pendiente igual a $-R_i$, según se representa en la fig. 5.9b. De la figura se desprende que la fuente se comporta como una fuente de voltaje constante, con un voltaje de salida V_o próximo al del generador V_i , solamente cuando el valor de R_i es pequeño o bien cuando el consumo de corriente es pequeño (de hecho ha de cumplirse que $IR_i \ll V_i$). Evidentemente, si V_i no es constante, caso frecuente, V_o tampoco es constante.

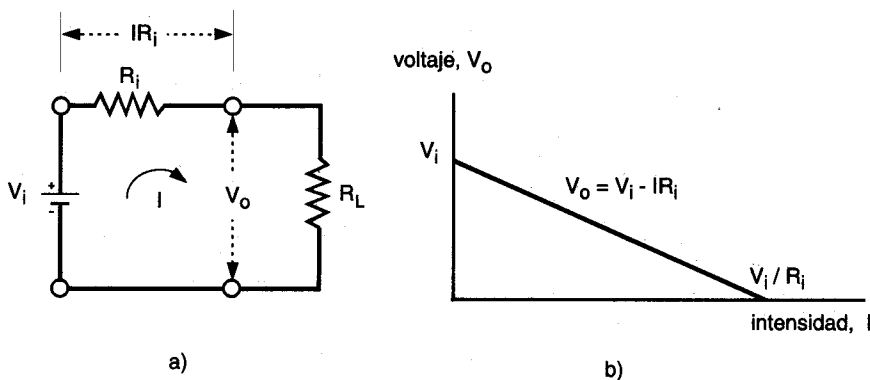


Fig. 5.9. a) Fuente de alimentación de voltaje para corriente continua formado por un generador de voltaje, V_i , con una resistencia interna R_i . b) Variación del voltaje de salida de la fuente, V_o , con la corriente I en el circuito externo.

En fuentes de alimentación de continua, una forma de estabilizar el voltaje de salida y de ampliar el rango de funcionamiento para corrientes elevadas (es decir para resistencias de carga pequeñas) consiste en la utilización de un diodo polarizado en inverso conectado en paralelo a la salida de la fuente (fig. 5.10a). Los diodos empleados para ejecutar este tipo de función se les denominan diodos Zener, ya que trabajan en el régimen de avalancha o ruptura. El fundamento de los diodos Zener ha sido ya descrito en un capítulo anterior (sec. 3.3.4). Como es sabido, cuando el diodo se polariza con un voltaje inverso superior o igual al voltaje de ruptura del diodo, V_z , la corriente I_z a través del diodo aumenta abruptamente (fig. 5.10b).

Veamos el modo de operación de un diodo Zener como regulador de voltaje. Sea la fuente de voltaje de la fig. 5.10a, formada por un generador de voltaje constante, V_i , una resistencia externa, R , en serie con el generador y un diodo Zener polarizado en inverso, el cual está conectado entre los terminales de salida. La resistencia R puede incluir la propia resistencia interna del generador. Como veremos más adelante, para que el diodo funcione como estabilizador del voltaje de salida de la fuente, se ha de cumplir que $V_z < V_i$.

Para determinar el punto de funcionamiento de la fuente, consideremos primero el caso extremo indicado en la fig. 5.10a en el que no existe resistencia de carga ($R_L = \infty$), lo cual implica que la corriente en el circuito externo es cero, es decir $I = 0$. En estas condiciones toda la corriente que suministra el generador se consume a través del diodo, y dado que $V_z < V_i$, la caída de potencial en el diodo se "autoajusta" al valor V_z para dar lugar a una corriente, que denominaremos I_z' , correspondiente a un punto determinado Q' en la región de ruptura de la

curva I_z - V del diodo (fig. 5.10b). Ahora bien, si la caída de potencial en el diodo es fija e igual a V_z entonces **la caída de potencial en la resistencia R ha de ser constante con un valor dado por $V_i - V_z$** . La corriente a través de esta resistencia (que es la misma que circula por el diodo) vendrá dada por:

$$I_z' = \frac{V_i - V_z}{R} \quad [5.2]$$

Es importante hacer notar que si por cualquier circunstancia la tensión del generador, V_i , varía la corriente I_z' variará también en la misma proporción, manteniéndose en todo caso el valor de V_z constante. Esto quiere decir que el potencial de salida de la fuente, V_o , se mantiene igualmente constante, ya que $V_o = V_z$.

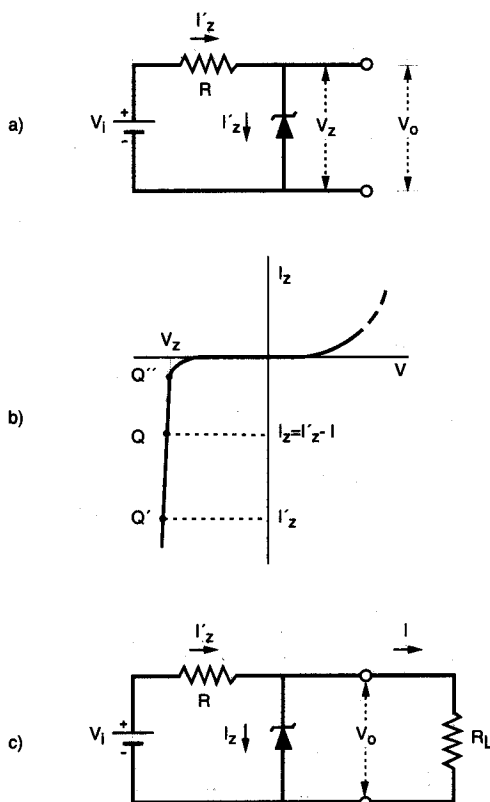


Fig. 5.10. a) Fuente de alimentación de continua estabilizada con un diodo Zener conectado a la salida de la fuente. b) Curva I_z - V característica del diodo en la región inversa. c) Esquema de la fuente de alimentación anterior con una resistencia de carga, R_L , conectada en los terminales de salida.

Cuando se añade una resistencia de consumo, R_L , en el circuito externo de la fuente consumiendo una cierta intensidad I ($I = V_o / R_L$), la caída de tensión a través del diodo en principio no tiene porqué variar siempre que R_L se mantenga por encima de ciertos límites. La caída de tensión en el diodo se "autoajusta" de nuevo en el valor V_z , por lo que la diferencia de potencial en la resistencia R seguirá siendo $V_i - V_z$. La corriente total a través de la resistencia R debe ser: $I + I_z$, donde I_z es la corriente que circula ahora a través del diodo (fig. 5.10c). El valor de la corriente a través de R vendrá dado por una ecuación similar a la ec. [5.2], es decir,

$$I_z + I = \frac{V_i - V_z}{R}$$

Esta ecuación nos permite obtener el nuevo valor de la corriente a través del diodo, esto es:

$$I_z = \frac{V_i - V_z}{R} - I \quad [5.3]$$

Este resultado indica que cuando existe una resistencia de consumo conectada a la fuente el punto de funcionamiento del diodo se traslada en la curva $I_z - V$ en la cantidad I , respecto el caso anterior (punto Q en la fig. 5.10b). De la expresiones [5.2] y [5.3] se desprenden además que:

$$I_z = I_z' - I \quad [5.4]$$

Esto quiere decir que, al conectar la resistencia R_L , la corriente en el diodo disminuye en la misma cantidad que aumenta la corriente en el circuito de consumo, con objeto de mantener la caída de tensión en la resistencia R constante. Así pues, **el efecto estabilizador del diodo Zener es debido a que las posibles variaciones en la corriente I en la resistencia externa originan en el diodo una variación de la corriente prácticamente igual pero en sentido opuesto, manteniéndose en todo caso la corriente a través de la resistencia R constante.** Asimismo, el potencial en el diodo, V_z , y el potencial de salida, V_o , se mantienen también constantes. Argumentos similares permiten demostrar que el potencial de salida también se mantiene constante incluso frente a posibles variaciones de la tensión del generador, siempre dentro de ciertos límites.

Hay que señalar, sin embargo, que, cuando el valor de R_L disminuye, el valor de I no puede aumentar indefinidamente, ya que I tiene un límite máximo, $I_{\max} = I_z'$, impuesto por la resistencia R cuando la fuente de voltaje trabaja en circuito abierto (ec. 5.2). Cuando I se acerca al límite máximo, la corriente en el propio diodo se hace muy pequeña, según indica la ec. [5.4], y el punto de funcionamiento del diodo se traslada a Q'' situado en el punto de inflexión la curva $I_z - V$ (fig. 5.10b). Cualquier aumento ulterior de la corriente I daría lugar a un desplazamiento del punto Q en la región horizontal de la curva $I_z - V$, con la correspondiente disminución del voltaje a través del diodo, V_z . La caída de tensión en la resistencia R

ya no sería constante y por consiguiente, el voltaje de salida de la fuente, $V_o (= V_z)$, también disminuiría.

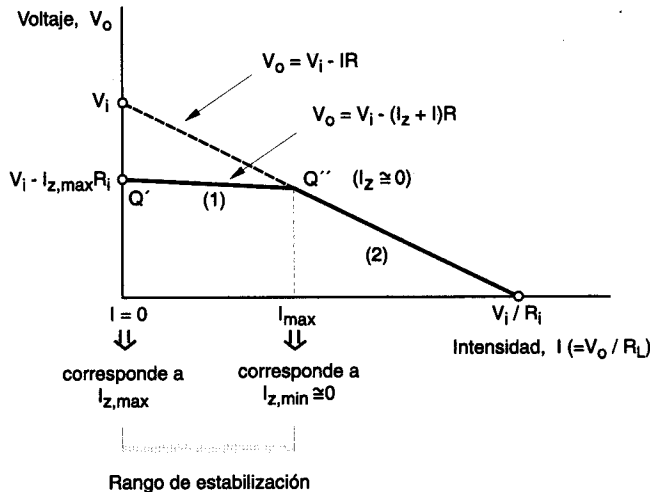


Fig. 5.11. Variación de la tensión de salida de la fuente de alimentación estabilizada (curva 1) y sin estabilizar (curva 2).

Así pues, dentro del rango de variación de la corriente en la resistencia de carga, desde $I = 0$ hasta $I \leq I_{max}$, el voltaje de salida se mantiene prácticamente constante en el valor $V_o = V_z$ (curva 1, fig. 5.11). En cambio, cuando la corriente I sobrepasa el valor $I = I_{max}$ el punto Q de funcionamiento del diodo se sale de la región de ruptura. En estas condiciones se puede decir que el diodo no tiene ningún efecto sobre el circuito, ya que la corriente a través de él es muy pequeña. Como consecuencia de ello, el voltaje V_o tiene entonces una variación con la corriente I similar a la de una fuente sin estabilizar (ec. 5.1), es decir: $V_o = V_i - IR$ (curva 2 de la fig. 5.11).

Curiosamente la ganancia en estabilidad de la fuente se consigue a costa de una pérdida sensible en la tensión de salida V_o respecto a la tensión del generador V_i (compárese en la fig. 5.11 la curva 1 con la de trazos, correspondiente al mismo generador funcionando sin diodo Zener). Por ello, si se pretende diseñar una fuente estabilizada con un voltaje de salida V_o , es preciso elegir el generador con un valor V_i más elevado que V_o . Al mismo tiempo el diodo Zener ha de tener un voltaje de ruptura, V_z , muy próximo al valor, V_o , que se desea en la salida. Normalmente el valor de V_z en un diodo Zener está determinado por la concentración de

impurezas a uno y otro lado de la unión. Esto significa que siempre es posible diseñar y fabricar diodos Zener con un voltaje de ruptura apropiado.

Asimismo, el valor $I_z' = I_{\max}$, permite determinar el rango de corriente en el circuito de consumo en el cual la fuente queda estabilizada. Según hemos visto, este rango comprende desde $I = 0$, para el cual I_z alcanzaría su valor máximo, I_z' , hasta un valor de $I = I_{\max}$, correspondiente a $I_z = 0$ (véase fig. 5.10b). El valor de I_z' depende de la resistencia R elegida en el circuito de estabilización (ec. 5.2), por lo que normalmente se toma R de tal manera que dé el valor más alto posible de I_z' compatible con las propias características de diseño del diodo. En particular, la corriente máxima a través del diodo está determinada por la capacidad de disipación de calor en el dispositivo, y su valor (junto con el de V_z) suele venir incluido en la especificaciones dadas por el fabricante.

5.4. DIODOS ESPECIALES (*)

Esta categoría de diodos especiales abarca un conjunto amplio de diodos que poseen un diseño especial y que basan su acción, más que en el efecto rectificador, en la asimetría de la propia unión p-n.

5.4.1. Diodos inversos

Según vimos en el capítulo tercero, el aumento de la concentración de impurezas en ambos lados de la unión p-n da lugar a un aumento del potencial de contacto, V_o , y a una disminución de la anchura de la región de agotamiento. La barrera de potencial en la unión p-n se hace en estas condiciones cada vez más abrupta, lo cual favorece el mecanismo de ruptura por efecto túnel cuando se alcanza una tensión crítica en polarización inversa. Como ya sabemos, en este proceso, debido al elevado campo eléctrico que existe en la unión, los electrones de la banda de valencia rompen su enlace con los átomos de la red y pasan por efecto túnel a la banda de conducción. El electrón y el hueco generados en la banda de conducción y en la de valencia, respectivamente, se mueven entonces en direcciones opuestas junto con los portadores mayoritarios hacia los electrodos (véase fig. 3.7).

Si la concentración de impurezas se hace suficientemente elevada se puede conseguir que la tensión crítica necesaria para la producción de pares electrón-hueco por efecto túnel sea prácticamente cero. Lo que ocurre entonces es que el aumento de la corriente, típico de un diodo en polarización inversa, aparece en una región cercana al origen en la curva I-V (fig. 5.12, parte superior). Los diodos fabricados con estas características se les denomina *diodos inversos*, porque en la región próxima al origen la curva I-V recuerda la de un diodo funcionando al revés, es decir, con más corriente en polarización inversa que en directa. Estas

características hacen que los diodos inversos sean muy adecuados para la rectificación y detección de señales muy débiles, inferiores al voltaje umbral, V_p , de un diodo, ya que un diodo normal solamente conduce de modo apreciable cuando la tensión aplicada a sus extremos es superior al voltaje umbral (véase por ejemplo la fig. 5.1).

En la fig. 5.12a se presenta un esquema del diagrama de bandas de energía de un diodo inverso sin polarización. Debido a los elevados dopajes empleados en la fabricación del diodo, el fondo de la banda de conducción del lado n se encuentra prácticamente a la misma altura que el tope de la banda de valencia del lado p. La aplicación de una tensión pequeña en inversa (tramo b en la curva I - V) hace que el tope de esta banda sobrepase al fondo de la banda de conducción del lado n, dando lugar a la transición de electrones de una banda a otra

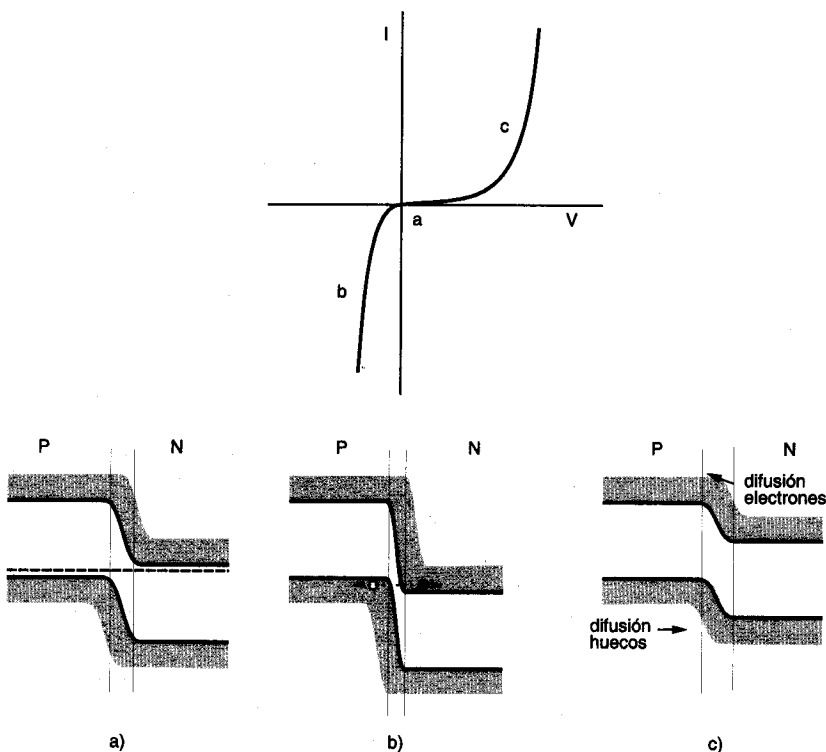


Fig. 5.12. Parte superior: Representación de la característica I-V de un diodo inverso. Parte inferior: a) Esquema del diagrama de bandas de energía de un diodo inverso sin polarización externa. b) Con polarización en inverso. c) Con polarización en directo.

por efecto túnel. En el diagrama de energías esta transición se verifica en horizontal (según se indica en la fig. 5.12b), ya que la energía potencial del electrón no varía por efecto de la transición. En esta región de voltajes donde predomina el efecto túnel la corriente a través del diodo es muy elevada. En cambio, en polarización directa (tramo c) el movimiento de los portadores se realiza como en un diodo normal, por lo que la corriente tiene un aumento exponencial de acuerdo con la ley del diodo (fig. 5.12c).

Cualquiera que sea la polarización aplicada al diodo inverso, la conducción se verifica siempre mediante los portadores mayoritarios (tanto si se polariza el diodo en directo como en inverso), los cuales responden de una manera muy rápida, en un tiempo inferior al nanosegundo, a las variaciones del campo eléctrico aplicado (recuérdese la discusión de la sección 3.5.3). Por esta razón, los diodos inversos pueden ser utilizados a frecuencias muy elevadas, en la región de los gigahercios.

5.4.2. Diodos túnel

Cuando la concentración de impurezas añadidas a un semiconductor alcanza un valor próximo o superior a la densidad efectiva de estados, es decir, del orden de 10^{19} átomos/cm³, se dice que el semiconductor es *degenerado* (ver sección 2.2.2). En estas condiciones, el nivel de Fermi del semiconductor puede estar dentro de la banda de conducción o de valencia, según sea el caso, y además los niveles de energía de las impurezas se transforman en bandas. La consecuencia más importante de ello es una disminución apreciable de la banda de energía prohibida del semiconductor.

Al poner en contacto dos semiconductores degenerados de carácter p y n, la igualación de los niveles de Fermi implica que el fondo de la banda de conducción del lado n puede estar por debajo del tope de la banda de valencia en el lado p. La barrera de la unión es además muy abrupta por lo que el paso de electrones de una banda a otra por efecto túnel puede ocurrir incluso cuando la unión p-n está sin polarizar.

La forma de la curva I-V para los diodos preparados de esta manera, denominados *diodos túnel*, viene dada en la fig. 5.13 (parte superior). En esta curva se observa que existe una región en la cual la corriente disminuye al aumentar el voltaje, es decir, **el diodo presenta una resistencia dinámica negativa**. Inicialmente se consideró esta variación como un efecto anormal para un diodo, y fue Esaki en 1958 el primero en encontrar una explicación basada en fenómenos cuánticos. Para entender mejor el origen de este efecto, en la parte inferior de la fig. 5.13 se da la posición relativa de las bandas de energía de un diodo túnel en diferentes estados de polarización. Cuando la tensión de polarización es cero, los electrones pasan por efecto túnel de la banda de valencia a la de conducción o en sentido inverso. El único requisito es que existan estados vacantes para esa energía a los cuales se pueda trasladar el electrón (diagrama b, de la fig. 5.13). Para tensiones de polarización inversas, existe un desnivel entre los niveles de Fermi de la banda de valencia del lado p y de la banda de conducción del lado n,

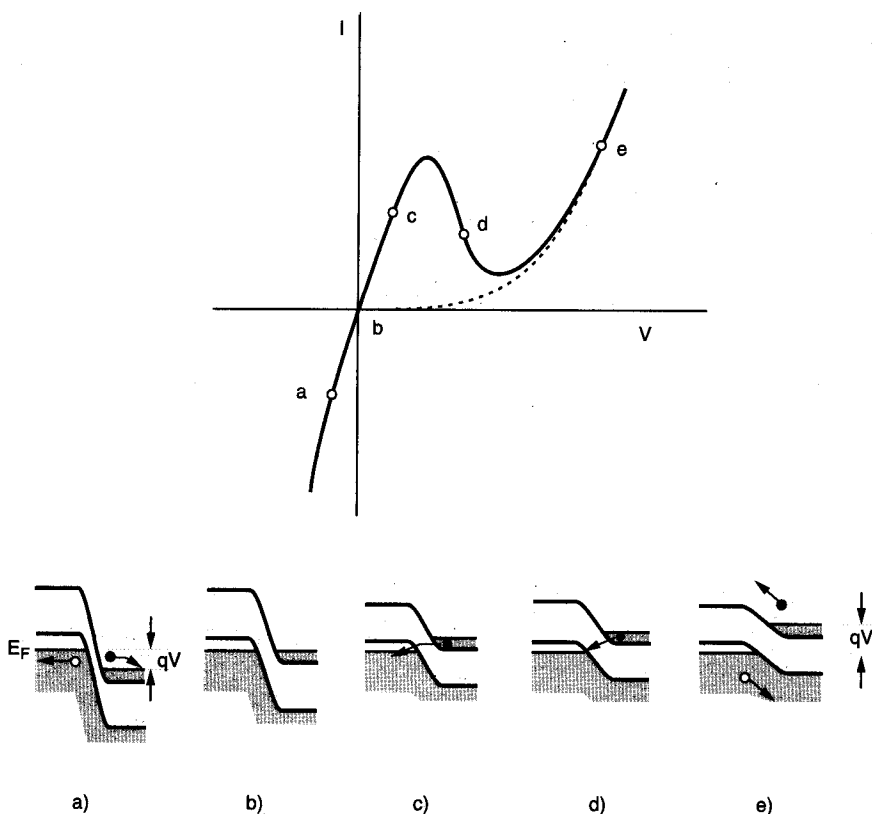


Fig. 5.13. *Parte superior: Característica I-V de un diodo túnel (la curva a trazos representa la corriente debida a la difusión de portadores mayoritarios). Parte inferior: Diagrama de bandas de energía en un diodo túnel en diferentes estados de polarización: a) Polarización inversa. b) Sin polarización externa. c) Polarización en directo con un voltaje aplicado inferior al crítico. d) idem. con un voltaje aplicado superior al crítico, y e) idem. con un voltaje muy alto.*

por lo que la probabilidad de salto por efecto túnel desde la banda de valencia a la de conducción es mayor, ya que el número de estados vacantes para el electrón es también mayor. Por tanto, la corriente en sentido inverso aumenta al aumentar el voltaje inverso (diagrama a). En polarización directa, si el voltaje aplicado es inferior a un valor crítico, existe también efecto túnel de electrones en sentido opuesto al anterior, es decir, desde un estado ocupado en la banda de conducción del lado n a un estado vacío de la banda de valencia del lado p (diagrama c). En esta situación la corriente aumenta con el voltaje. Cuando se aplican voltajes superiores al crítico, el fondo de la banda de conducción puede sobrepasar al borde superior de la

banda de valencia. En esta región de voltajes, disminuye el número de estados ocupados en la banda de conducción susceptibles de soltar un electrón que pueda "tunear" hacia la banda de valencia (diagrama d) y la corriente disminuye con el voltaje aplicado (zona de resistencia dinámica negativa). Mayores incrementos del voltaje reducen aún más la "corriente túnel". Sin embargo, a partir de un cierto voltaje la corriente de difusión normal en un diodo empieza a predominar sobre la de efecto túnel. La corriente aumenta de nuevo con el voltaje siguiendo la ley del diodo (diagrama e).

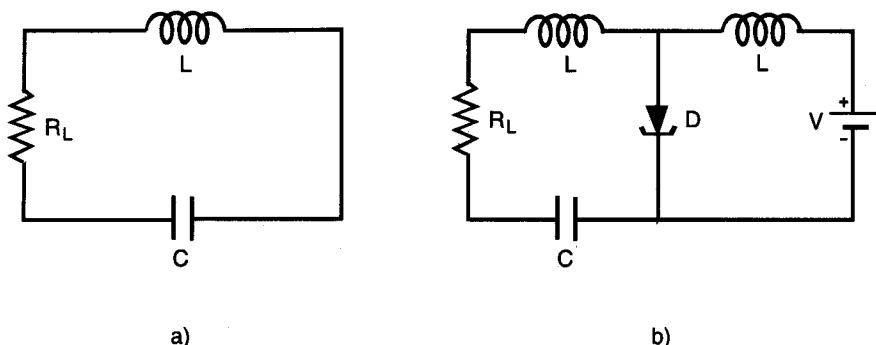


Fig. 5.14. a) Circuito oscilador simple formado por un condensador y una bobina en serie. b) Circuito oscilador incluyendo un diodo túnel, D, para compensar las pérdidas asociadas a la resistencia del circuito (nótese el símbolo empleado para representar el diodo túnel).

La región de resistencia negativa del diodo túnel es muy útil para su utilización en circuitos osciladores, tipo LC. En estos circuitos la onda alterna inducida se amortigua debido a las pérdidas inevitables en las resistencias inherentes a la bobina y al cableado. La introducción de un diodo túnel, polarizado en la zona de resistencia negativa permite compensar estas pérdidas. En la fig. 5.14 se da un esquema de un circuito oscilador tipo LC simple (fig. 5.14a) y del mismo circuito incluyendo un diodo túnel, D, para compensar las pérdidas asociadas a la resistencia de carga, R_L y al resto del circuito (fig. 5.14b). El diodo está acompañado de una fuente de alimentación que polariza el diodo y es en definitiva la que suministra la energía perdida. Debido a que en el efecto túnel sólo participan portadores mayoritarios, la velocidad de respuesta de los diodos a variaciones de potencial es muy rápida. Por esta razón la frecuencia de oscilación puede sobrepasar el gigahercio. Sin embargo, la amplitud de oscilación es pequeña, ya que está limitada a la región de voltajes donde se presenta la resistencia negativa.

5.4.3. Diodos de capacidad variable: varactores

En los diodos semiconductores, la variación de la capacidad de la unión con el voltaje de polarización (apartado 3.5) encuentra numerosas aplicaciones en circuitos electrónicos, tales como en circuitos de control de frecuencia para sintonía, en circuitos de amplificación paramétrica, etc. Según se vió en el capítulo tercero, la principal contribución a la capacidad de un diodo proviene de dos factores: i) la capacidad, C_s , debida a la carga espacial localizada en la región de agotamiento, y ii) la denominada capacidad de difusión, C_d , asociada a los portadores minoritarios acumulados en las regiones neutras. La capacidad C_s está determinada por el voltaje de polarización aplicado al diodo (ec. 3.48), mientras que la capacidad C_d depende sobre todo de la corriente que pasa a través de la unión (ec. 3.51). Por esta razón, esta última contribución es dominante sobre todo para voltajes de polarización en directo, mientras que en polarización inversa la contribución más importante es debida a la capacidad asociada a la carga espacial. Así pues, es en la región de voltajes negativos (polarización inversa) donde los diodos pueden ser utilizados como dispositivos de capacidad variable, controlada por el voltaje aplicado al diodo. A estos dispositivos se les conoce con el nombre de varactores¹.

De acuerdo con la ec. (3.48), la dependencia de la capacidad con el voltaje para una unión abrupta p^+-n (es decir, $N_a \gg N_d$) viene dada por una función del tipo $C_s \propto V^{-1/2}$, siendo V el voltaje aplicado en sentido inverso. A este tipo de unión también se le denomina *unilateral*. Muy a menudo interesa tener otro tipo de dependencia de la capacidad con el voltaje. Así, por ejemplo, en una unión también unilateral, con una variación lineal de la concen-

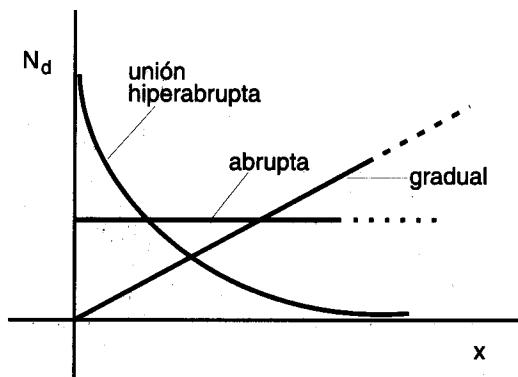


Fig. 5.15. Variación de la concentración de impurezas con la distancia a la unión, x , en un unión $p-n$ unilateral para tres tipos diferentes de dopaje en el lado n .

¹ Nota: El nombre varactor procede del idioma inglés a través de la contracción de los términos: variable-reactor.

tracción de donadores con el espesor, *unión gradual* (fig. 5.15), se encuentra que la relación entre la capacidad y el voltaje inverso es del tipo $C_s \propto V^{-1/3}$. Ambas dependencias obedecen a una ley más general que tiene la forma:

$$C \propto V^{-n} \quad [5.5]$$

con $n = 1/2$ para la unión abrupta y $n = 1/3$ para la unión gradual. Los diodos varactores requieren en muchas aplicaciones una dependencia de la capacidad con el voltaje lo más fuerte posible. En estos casos se recurre a la denominada *unión hiperabrupta*. En este tipo de unión la concentración de impurezas es muy elevada en la superficie de la unión y disminuye gradualmente a medida que aumenta la distancia x hacia el interior (ver fig. 5.15). La dependencia de la capacidad con el voltaje es entonces muy acusada, con $n = 2$ en la ecuación anterior. Esta fuerte dependencia es explicable si se tiene en cuenta que la variación de la capacidad procede de la variación del espesor de la región de carga espacial con el voltaje aplicado. Por las especiales características de la unión hiperabrupta, la variación del espesor produce cambios muy grandes en la carga contenida en la región de agotamiento.

5.4.4. Diodos p-i-n

La introducción de una capa intrínseca entre dos semiconductores de tipo p y n, *diodos p-i-n*, tiene indudables ventajas en diversas aplicaciones de los diodos, entre ellas la posibilidad de soportar potenciales de ruptura mucho más elevados. Efectivamente, la presencia de la región intrínseca (región i), hace que en los diodos p-i-n, la distribución de la carga espacial a lo largo de la unión cuando los semiconductores se encuentran en equilibrio sin tensión aplicada sea según se indica en la fig. 5.16a. De acuerdo con esta figura, la carga espacial positiva

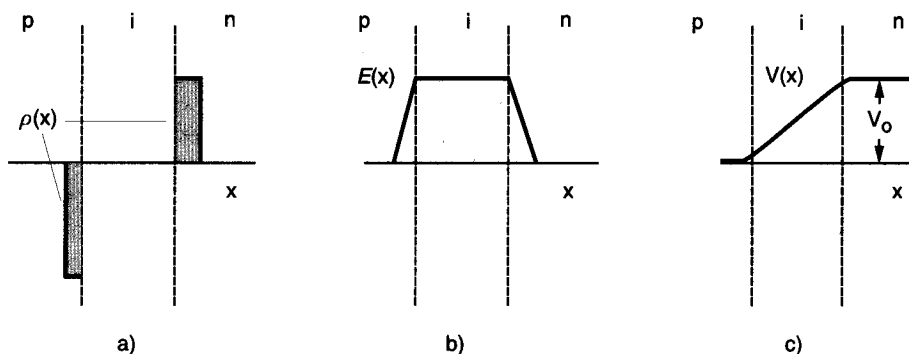


Fig. 5.16. Variación a) de la carga espacial, $\rho(x)$, b) del campo eléctrico, $E(x)$, y c) del potencial $V(x)$ a lo largo de la unión de un diodo p-i-n en equilibrio, es decir, sin tensión aplicada.

y negativa se sitúa a ambos lados de la región intrínseca. Ello es debido al trasvase de huecos de la región p a la región intrínseca y de electrones de la región n a la intrínseca. Asimismo, desde la región intrínseca existe un trasvase de electrones a la región p, y de huecos a la región n. Un análisis detallado a través de la ecuación de Poisson permite obtener la variación del campo eléctrico E y el potencial de contacto V_0 . Ambas magnitudes han sido también representadas en las figs. 5.16b y 5.16c. Obsérvese que en la región intrínseca el campo eléctrico E asociado a la unión es constante y por tanto la variación del potencial es lineal. Se puede decir que los diodos p-i-n son similares a un diodo p-n, pero con una región de agotamiento muy amplia. En cualquier caso el campo eléctrico de la unión en los diodos p-i-n es menor que en los diodos p-n. Este hecho hace que los diodos p-i-n puedan soportar voltajes en inverso elevados sin alcanzar la región de ruptura. La baja capacidad asociada a la región de agotamiento permite que los diodos p-i-n puedan ser utilizados en circuitos osciladores de frecuencias elevadas, compitiendo incluso en algunas aplicaciones con los diodos túnel.

5.5. DISPOSITIVOS OPTOELECTRONICOS

La interacción de la radiación electromagnética con los átomos de un material semiconductor produce una gran variedad de fenómenos que a su vez han sido aprovechados ventajosamente en la fabricación de diversos dispositivos denominados *optoelectrónicos*, entre ellos los diodos detectores de radiación, células solares, diodos emisores de luz, láseres, etc. Antes de entrar en la descripción detallada de estos dispositivos, conviene hacer un repaso general de la interacción de la luz con los átomos de un medio cualquiera.

Entre los fenómenos más importantes producidos por la interacción luz-materia tenemos la transición radiativa o emisión y la absorción óptica. En la fig. 5.17 viene representado cada uno de estos procesos en un esquema de niveles discretos de energía. En la *absorción óptica* un electrón en un estado inicial de energía E_1 pasa por efecto de la radiación a un estado final de energía E_2 (fig. 5.17a). En este proceso de absorción de energía entre niveles discretos se ha de cumplir que la diferencia de energías entre el estado final y el inicial sea exactamente igual a la energía de la radiación, esto es: $E_2 - E_1 = h\nu$. Una ecuación similar se cumple para las transiciones en sentido inverso -*procesos de emisión*-, en las cuales un electrón pasa desde un estado inicial E_2 de energía elevada a otro E_1 de energía más baja, en este caso con emisión de fotones de energía $h\nu$.

Los procesos de emisión se suelen subdividir a su vez en los denominados espontáneos y estimulados. En los procesos de *emisión espontánea* un electrón que se encuentra en un estado excitado E_2 pasa después de un pequeño período de tiempo al estado fundamental E_1 sin que exista ningún estímulo externo, emitiendo un fotón de energía $h\nu$ (fig. 5.17b). En cambio, en la *emisión estimulada* el electrón realiza la emisión desde el estado excitado solamente cuando el átomo recibe un fotón cuya energía $h\nu$ coincide con la diferencia $E_2 - E_1$. Por tanto, el fotón emitido y el incidente tienen exactamente la misma frecuencia y además oscilan con la misma fase (fig. 5.17c). La radiación emitida por el conjunto de átomos del material se-

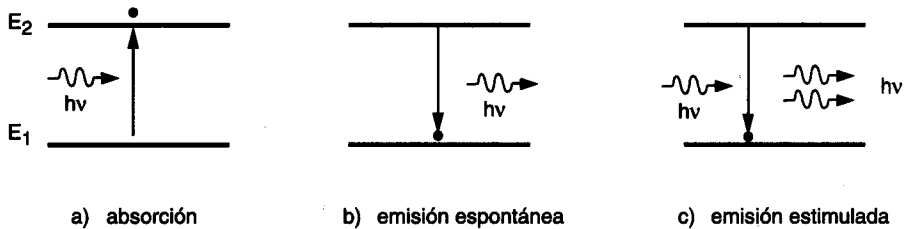


Fig. 5.17. *Procesos básicos de emisión y absorción de radiación electromagnética entre dos niveles energéticos de un átomo.*

rá por tanto monocromática y coherente (todos los fotones oscilan con la misma frecuencia y fase). Se puede demostrar que **para que exista emisión estimulada el estado de ocupación de los niveles excitados tiene que ser mayor que el de los niveles fundamentales**. A esta condición se la denomina *inversión de poblaciones*, ya que en un material en equilibrio térmico ocurre lo contrario, es decir, el nivel fundamental está más poblado que los niveles excitados. Además se requiere también que exista una elevada densidad de energía de radiación $h\nu$ para que se produzca emisión estimulada. Como veremos después existen diversos procedimientos para que se cumplan estas dos condiciones.

Cuando se trata de un material semiconductor la distribución de los electrones en bandas de energía, en lugar de niveles discretos, hace que en los procesos de absorción la energía del fotón necesaria para excitar un electrón desde la banda de valencia a la de conducción cumpla la condición:

$$h\nu \geq E_c - E_v = E_g \quad [5.6]$$

siendo E_g la energía de la banda prohibida. Ello es debido a que el electrón se puede trasladar desde un nivel de energía inferior al tope de la banda de valencia a otro nivel con energía superior a la del fondo de la banda de conducción. Cuando la transición se realiza desde o hacia estados asociados a impurezas o a otros defectos del material con estados energéticos en la banda prohibida, la energía de la transición es evidentemente menor. Nótese que los procesos de absorción y emisión se corresponden con los de excitación y desexcitación de electrones en un semiconductor, ya tratados en los capítulos 1 y 2.

En los procesos de absorción óptica la luz se atenúa a medida que atraviesa el medio. Dado que la cantidad de luz absorbida en un elemento infinitesimal de volumen es siempre

proporcional a la intensidad de luz existente en ese punto, la disminución de la intensidad de la luz, I , con la distancia x recorrida desde la superficie seguirá una ley exponencial del tipo:

$$I = I_0 \exp(-\alpha x) \quad [5.7]$$

siendo I_0 la intensidad de luz incidente en la superficie del material y α el *coeficiente de absorción*. Este coeficiente está relacionado con la proporción de luz que es absorbida por unidad de longitud, de forma que cuanto mayor sea su valor mayor será la atenuación de luz para un espesor dado de material. El coeficiente α depende mucho de las características del medio y sobre todo, en el caso de semiconductores, de la longitud de onda λ de la radiación ($\lambda = c/v$, siendo c la velocidad de la luz en el vacío). La longitud de onda crítica, λ_c , por debajo de la cual la energía de la radiación es suficiente para excitar electrones desde la banda de valencia a la de conducción vendrá dada por:

$$\lambda_c = \frac{c}{v} = \frac{hc}{hv} = \frac{1.24}{E_g} \mu\text{m} \quad [5.8]$$

con E_g expresado en electrón-voltio. Si la energía de la radiación es tal que $\lambda < \lambda_c$, el coeficiente de absorción será elevado. En cambio, cuando λ se acerca o se hace mayor que λ_c , el coeficiente de absorción disminuye. En la fig. 5.18 se presenta la variación experimental de α en función de λ para algunos semiconductores típicos. Obsérvese que la disminución de α se

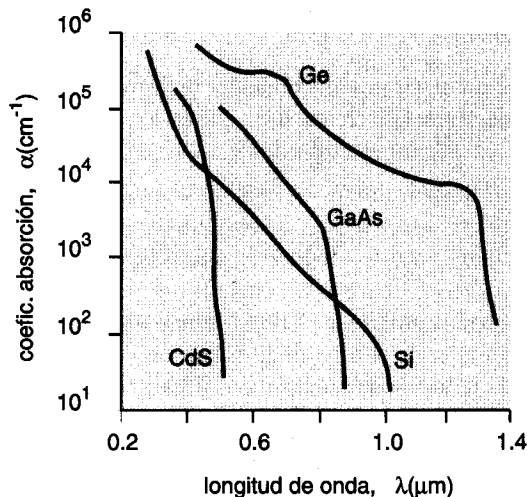


Fig. 5.18. Variación del coeficiente de absorción en función de la longitud de onda de la radiación incidente, para algunos semiconductores típicos.

hace muy abrupta cuando se alcanza el valor crítico que corresponde con el comienzo de la transición de banda a banda (borde de absorción). De hecho, la técnica de obtención de estas curvas, denominada *espectroscopia óptica*, constituye un buen método de determinación de λ_c , y por tanto de la energía de la banda prohibida de un semiconductor (ec. 5.8).

5.5.1. Fotoconductores (*)

Las transiciones de banda a banda (intrínsecas) o las que se hacen involucrando niveles de impurezas (extrínsecas) en los procesos de absorción óptica en un semiconductor dan lugar a un incremento de la concentración de portadores en la banda de valencia o de conducción del semiconductor. Este fenómeno produce un aumento de la conductividad del material, y se utiliza para detectar y medir la intensidad de la radiación. Así, en los dispositivos fotoconductores se utiliza simplemente un semiconductor muy sensible a la radiación luminosa en una región de longitud de onda determinada, el cual es sometido a un voltaje V mediante dos electrodos aplicados en sus extremos formando un contacto óhmico (fig. 5.19a).

Consideremos el caso de un semiconductor intrínseco, en el cual la concentración de portadores es muy baja. En ausencia de radiación (oscuridad) su conductividad ha de ser baja, y por tanto la corriente, I , medida cuando se aplica un voltaje V en los extremos del semiconductor también será muy pequeña. En cambio, cuando el semiconductor se ilumina con una radiación de longitud de onda adecuada, la concentración de portadores aumenta sensiblemente sobre la que corresponde al equilibrio térmico, debido a la generación de pares electrón-hueco, según se muestra en el esquema de bandas de energía de la fig. 5.19b. En estas condiciones la conductividad del material aumenta en proporción a la intensidad de la luz y puede alcanzar valores muy elevados. La corriente a través del material debe tener entonces un valor elevado.

Es importante destacar que **no todos los portadores fotogenerados contribuyen a la conducción, ya que una fracción importante de ellos se recombina antes de llegar al extremo correspondiente del semiconductor**. Se puede hacer un cálculo sencillo del incremento de corriente, ΔI_e , debida al exceso de electrones generado en la banda de conducción, Δn , a través de la ecuación:

$$\Delta I_e = q \mu_e (\Delta n) E S \quad [5.9]$$

siendo E el campo eléctrico aplicado, μ la movilidad de los electrones y S la sección transversal del fotoconductor. En condiciones de iluminación, el estado estacionario se alcanza cuando la velocidad de generación de portadores en todo el volumen del semiconductor, G , se iguala a la velocidad de recombinación, R , es decir, $R = G$. Según se estableció en el apartado

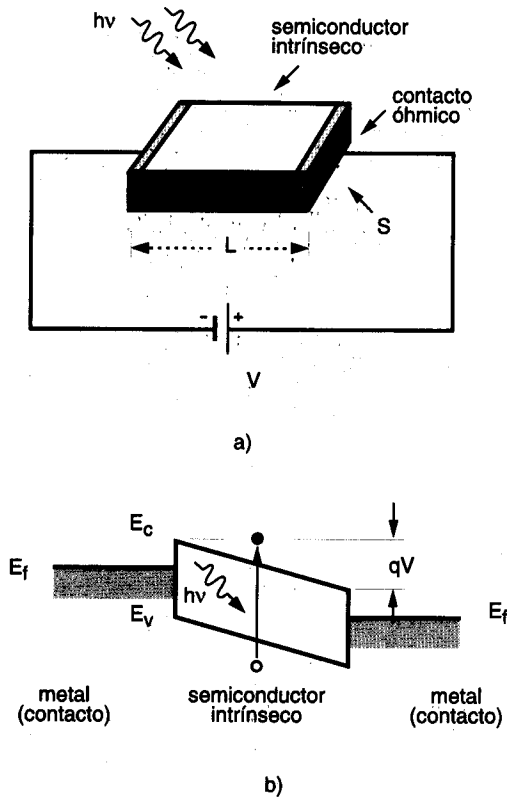


Fig. 5.19. a) Esquema del funcionamiento de un dispositivo detector de radiación utilizando un material fotoconductor. b) Esquema de bandas de energía del proceso de generación de pares electrón-hueco en el fotodetector sometido a una diferencia de potencial V .

2.6.1, para un semiconductor intrínseco en el cual existe un exceso de portadores, $\Delta n = \Delta p$, la velocidad de recombinación de los portadores vendrá dada por:

$$R = \frac{\Delta n}{\tau} = \frac{\Delta p}{\tau} \quad [5.10]$$

siendo τ el tiempo de vida media de los portadores fotogenerados. Por otra parte, en un semiconductor de longitud L en el que suponemos que el espesor es suficiente para que toda la

luz que incide sobre él sea absorbida en su interior, tendremos para la velocidad de generación de portadores en la banda de conducción:

$$G = \eta n_{\text{fot}} = \eta \frac{P_i / h\nu}{SL} \quad [5.11]$$

siendo n_{fot} el número de fotones incidentes en el semiconductor por unidad de volumen y de tiempo, y η la eficiencia de la conversión en la generación de portadores. El valor de n_{fot} se calcula en la última igualdad a través del cociente entre la potencia de la luz incidente, P_i , y la energía de la radiación, $h\nu$, dividido a su vez por el volumen del material.

Sabiendo que la velocidad de arrastre de los electrones por el campo eléctrico viene dada por: $v_e = \mu_e E$, las igualdades anteriores permiten escribir para la corriente de electrones fotogenerada entre los dos electrodos:

$$\Delta I_e = q v_e \eta \frac{P_i / h\nu}{L} \tau \quad [5.12]$$

Si tenemos en cuenta que el cociente $t_r = L/v_e$ representa el tiempo de tránsito de los electrones entre los dos electrodos, resulta para ΔI_e :

$$\Delta I_e = q \eta \frac{P_i}{h\nu} \frac{\tau}{t_r} \quad [5.13]$$

con una expresión similar para la corriente de huecos en la banda de valencia. En la ecuación anterior, el factor $q\eta(P_i/h\nu) = I_{\text{fot}}$ tiene dimensiones de corriente y representa la velocidad de generación de carga en el semiconductor. Por ello es considerado a menudo como la corriente primaria debida a los portadores fotogenerados. En función de este parámetro, se define el factor de ganancia del fotoconductor a través del cociente:

$$\frac{\Delta I}{I_{\text{fot}}} = \frac{\tau}{t_r} \quad [5.14]$$

En la ecuación anterior se ha prescindido del subíndice de ΔI , ya que el resultado es válido tanto para la corriente de electrones como para la corriente de huecos en el semiconductor, siempre que se utilice en cada caso el valor correspondiente de t_r . Con objeto de aumentar la ganancia del fotodetector interesa utilizar materiales en los cuales el tiempo de vida de los portadores generados, τ , sea lo mayor posible. Asimismo, el tiempo de tránsito de los portadores desde un electrodo a otro, t_r , ha de ser muy pequeño con objeto de que sean colectados antes de que se recombinen. Para ello es preciso utilizar semiconductores muy puros y libres de defectos. La longitud del dispositivo interesa que sea también muy pequeña y el campo eléctrico aplicado elevado. Para materiales fotoconductores típicos, como el sulfuro de cadmio

o el sulfuro de plomo, se ha encontrado para el cociente τ/t_r valores del orden de 10^6 . Los dispositivos fotoconductores se utilizan sobre todo en el infrarrojo (en la región de 8-14 μm de longitud de onda), donde no existen otras alternativas de detección. En estos casos es preciso utilizar semiconductores con una banda prohibida muy pequeña (el compuesto HgCdTe es uno de ellos), y refrigerarlos a temperaturas de 77 K mediante nitrógeno líquido con objeto de reducir la contribución de los portadores intrínsecos.

5.5.2. Diodos detectores de radiación: Fotodiodos

Muy a menudo se utilizan diodos de unión para mejorar la sensibilidad y la velocidad de respuesta de los detectores de radiación en la región óptica de más alta energía. Así cuando un diodo se ilumina con radiación de energía suficiente se crean pares electrón-hueco a ambos lados de la unión como consecuencia de la excitación de portadores desde la banda de valencia a la de conducción. Los portadores generados a uno y otro lado a distancias grandes de la unión no producen efectos apreciables en las características del diodo. En cambio, los pares electrón-hueco generados, bien sea dentro de la región de carga espacial o bien a una distancia de la unión menor que la correspondiente longitud de difusión, son arrastrados hacia el lado opuesto a causa del campo eléctrico presente en la unión. El exceso de carga creado en las regiones neutras a cada lado de la unión origina una diferencia de potencial, V_{oc} , que tiene la misma polaridad que el diodo, es decir, lado p positivo y lado n negativo. Este proceso de separación de los portadores se conoce como *efecto fotovoltaico*, y los diodos que emplean este efecto para detectar la presencia de radiación se denominan a su vez *fotodiodos* (fig. 5.20a).

Un diodo operando con un cierto voltaje aplicado, V , en presencia de radiación electromagnética capaz de excitar portadores a través de la banda prohibida dejará pasar una intensidad I dada por:

$$I = I_0 [\exp (qV / kT) - 1] - I_L \quad [5.15]$$

donde el primer término representa la corriente típica de un diodo, es decir la corriente en oscuridad (ec. 3.36) y el segundo término I_L representa la corriente debida a los portadores generados. Este término viene precedido del signo negativo porque el movimiento de estos portadores se verifica en la misma dirección que la de los portadores minoritarios, esto es, con el mismo sentido que la corriente en inversa del diodo. El valor de I_L puede calcularse a través de la ecuación:

$$I_L = q G S (L_e + L_h) \quad [5.16]$$

siendo G el número de portadores generados por unidad de volumen y de tiempo y S el área de la sección transversal del diodo. L_e y L_h representan las longitudes de difusión de electrones y huecos ya definidas en el capítulo segundo (apartado 2.6).

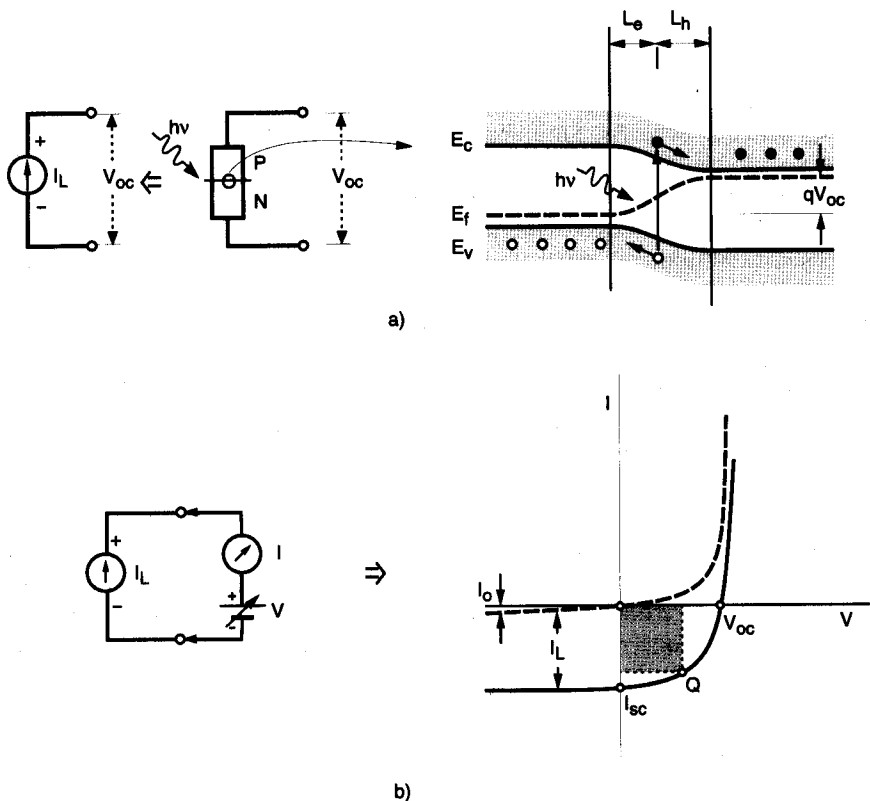


Fig. 5.20. a) Proceso de generación de pares electrón-hueco por absorción de luz en la región de carga espacial de una unión p-n (la figura de la izquierda representa el circuito equivalente de un fotodiodo formado por la fuente de corriente I_L). b) Medida de la curva característica I-V de un fotodiodo en iluminación. La línea a trazos representa la curva característica en oscuridad.

En la fig. 5.20b se ha representado la característica I-V de un fotodiodo bajo iluminación. Obsérvese que esta característica está desplazada en una cantidad I_L , prácticamente constante, respecto a la curva en la oscuridad. Los fotodiodos generalmente operan en el tercer cuadrante, es decir, con polarización negativa y con corriente también negativa, ya que en esta región la corriente es prácticamente independiente del voltaje y además proporcional a

la velocidad de generación de portadores (siempre que $I_L \gg I_0$). El dispositivo funciona entonces como detector del nivel de iluminación convirtiendo una señal óptica en señal eléctrica.

Con objeto de aumentar la velocidad de respuesta del fotodiodo normalmente se reduce la anchura de la región de agotamiento, ya que de esta manera el tiempo de tránsito de los portadores, t_r , es más pequeño. Sin embargo, por otra parte interesa también que la anchura de esta región sea lo mayor posible, ya que así la mayor parte de la radiación se absorbe en esta región. Por tanto si se quiere a la vez una alta velocidad de respuesta y una buena eficiencia en la conversión de la luz absorbida es preciso llegar a un compromiso. En este sentido se recurre muy a menudo a la utilización de diodos p-i-n, en los cuales la anchura de la región de agotamiento se puede variar con relativa facilidad, ya que ésta viene determinada fundamentalmente por la anchura de la capa intrínseca (fig. 5.21a).

En los *fotodiodos de avalancha*, el dispositivo se polariza en el régimen de avalancha de polarización inversa. De esta manera los pares electrón-hueco generados en la región de agotamiento pueden alcanzar un factor de multiplicación elevado mediante el proceso de avalancha, consiguiéndose así un factor de ganancia también elevado.

En la fig. 5.21b se presenta un diagrama esquemático del corte transversal de un fotodiodo tipo p-i-n. La región intrínseca está situada muy cerca de la superficie con objeto de aumentar al máximo la absorción de la radiación en esta región. El contacto metálico superior suele hacerse utilizando bien sea un material conductor transparente (el óxido de estaño o de indio puede ser adecuado) o bien una capa metálica muy fina en forma de rejilla dejando la

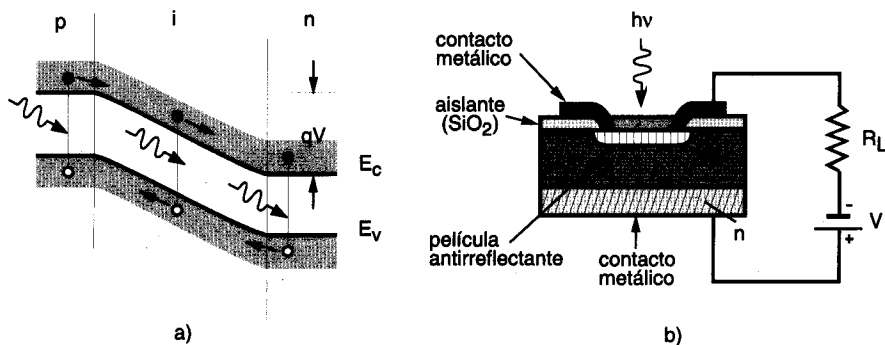


Fig. 5.21. a) Esquema del proceso de generación de portadores en un fotodiodo del tipo p-i-n. b) Sección transversal de un fotodiodo p-i-n (obsérvese la disposición del electrodo superior para permitir el paso de la radiación, y la presencia de un película antirreflectante).

máxima superficie del diodo expuesta a la radiación. El dispositivo lleva además una *capa antirreflectante* para disminuir al máximo las pérdidas por reflexión de la luz en la superficie del diodo. Estas capas están formadas por una película transparente de un material aislante (SiO_2 , Si_3N_4 , etc.), cuyo índice de refracción y espesor son los adecuados para evitar, mediante un fenómeno de interferencia, la reflexión de la luz.

5.5.3. Células solares

Si el diseño de un fotodiodo se optimiza para convertir la radiación solar en una corriente eléctrica susceptible de ser aprovechada en un circuito de consumo tenemos entonces una célula solar. En este caso el diodo opera en el cuarto cuadrante de la fig. 5.20b, es decir, suministrando un voltaje de salida y una corriente sin necesidad de polarización externa (punto Q en la característica I-V). Nótese que en esta región de operación el voltaje es positivo y la corriente es negativa, lo cual significa que la potencia disipada en el diodo es negativa, o dicho en otras palabras, el diodo entrega potencia al exterior. En cualquier caso, cuando la célula trabaja en un régimen de voltajes bajos su comportamiento se puede aproximar al de una fuente de corriente² de intensidad I_L , suministrando un voltaje variable, V , cuyo valor depende de la resistencia de carga conectada a la célula.

En la preparación de una célula solar se ha de poner especial atención en minimizar el valor de la resistencia serie de la célula, R_s . Esta resistencia procede de las resistencias asociadas a las regiones neutras y a los contactos de salida del diodo. En una célula solar típica, la presencia de una resistencia serie de sólo unos pocos ohmios puede reducir la potencia de salida en un factor 3 (véase problema A3). En la fig. 5.22a se presenta un esquema del circuito equivalente de una célula solar donde se ha incluido la resistencia serie, R_s , y la resistencia de carga, R_L . En este circuito la célula solar se ha representado por una fuente de corriente de valor I_L , con una resistencia en paralelo constituida por el propio diodo. La corriente $I_{\text{cel}} = I_0 [\exp(qV/kT) - 1]$, que circula a través del propio diodo, corresponde al primer término de la ecuación [5.15] y está originada por la corriente de mayoritarios debida al voltaje, V , suministrado por la propia célula. En el punto de operación, normalmente $I_{\text{cel}} < I_L$, por lo que la corriente I en el circuito externo tiene el mismo sentido que I_L .

Otro factor crítico a considerar en el diseño de las células solares es el punto de funcionamiento, Q, en la característica I-V. Según se comenta en el Apéndice A2, el punto de funcionamiento de un diodo está determinado por la intersección de la curva característica del diodo y la recta de carga. En una célula solar capaz de suministrar una corriente I sobre una resistencia de consumo, R_L , la recta de carga está determinada por la ecuación: $V = I R_L$, es decir, se corresponde con una recta que pasa por el origen y tiene una pendiente igual a $1/R_L$, tal como se indica en el diagrama de la fig. 5.22b. En esta figura se ha representado por co-

² Nota: En el Apéndice A3 se hace una descripción de las características de las fuentes de alimentación de voltaje y de corriente.

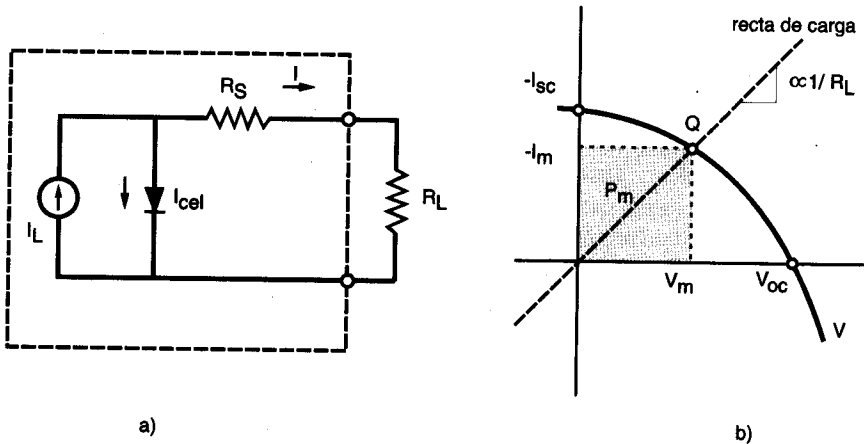


Fig. 5.22. a) Circuito equivalente de una célula solar considerada como una fuente de corriente. b) Representación de la curva característica I-V y del punto Q de funcionamiento de una célula solar en condiciones de máxima potencia de salida.

comodidad la parte negativa del eje de intensidades hacia arriba. La elección de R_L se hace siguiendo el criterio de que el punto Q óptimo es aquel en el cual la potencia de consumo, $P = VI$, es máxima. Esto quiere decir que el área del rectángulo inscrito dentro de la curva $I - V$, en el cuarto cuadrante debe ser máxima. Este área viene representada por el rectángulo rayado de la fig. 5.22b. Se puede demostrar que en el punto óptimo de operación, la potencia $P_m = V_m I_m$ es aproximadamente: $0.75 V_{oc} I_{sc}$, siendo V_{oc} e I_{sc} la tensión en circuito abierto y la corriente en cortocircuito, respectivamente. Interesa por tanto que el valor de V_{oc} de una célula sea lo más elevado posible, lo cual se consigue aumentando el dopaje de los semiconductores a cada lado de la unión. Sin embargo, el aumento del dopaje trae consigo una disminución de las longitudes de difusión, las cuales han de mantenerse también elevadas. Se hace necesario, también en este caso, llegar a una solución de compromiso.

Dado que el espectro solar tiene unas características muy específicas, mostrando un máximo de radiación en la región del visible (alrededor de $0.5 \mu m$), es preciso elegir en el diseño de la célula materiales semiconductores con una banda prohibida adecuada. Si la energía de la banda prohibida E_g es pequeña, gran parte de la energía solar, la de menor longitud de onda, λ , se absorbe directamente en forma de calor y no contribuye a la corriente de la célula. Al contrario, si E_g es elevada la radiación con un valor de λ mayor que el crítico, λ_c , no se absorbe.

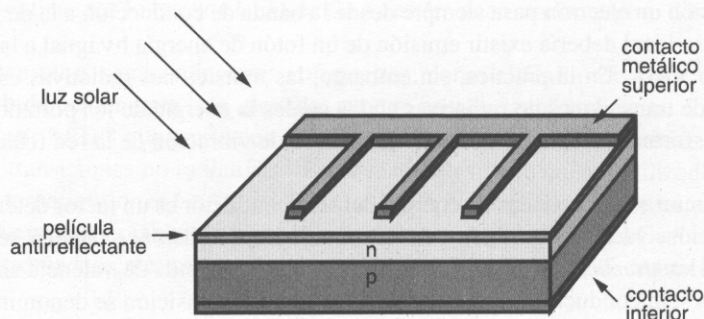


Fig. 5.23. Estructura de una célula solar típica.

Teniendo en cuenta todos los factores considerados es posible calcular la anchura óptima de la banda prohibida que ha de tener el semiconductor para coleccionar la máxima potencia de energía solar. Así se demuestra que el valor de E_g óptimo está situado entre 1.1 y 1.4 eV dependiendo de las condiciones de irradiación solar (terrestre o en el espacio). El *factor de eficiencia* en la captación, es decir, potencia suministrada dividida por la potencia incidente, puede ser entonces hasta del 30%. Los semiconductores que mejor se adaptan a estos valores de E_g son el Si y el GaAs, con valores de E_g de 1.12 y 1.43 eV, respectivamente (véase tabla 2.2). Otros materiales que tienen menos eficiencia pero que resultan muy atractivos para su utilización en células solares debido a su bajo coste de preparación son el a:Si (silicio amorfo) y el CdS, este último en forma policristalina.

En la fig. 5.23 se presenta un esquema de la estructura de una célula solar típica, en la que se incluye también una película antirreflectante para evitar las pérdidas por reflexión de la luz. El contacto eléctrico de salida en la parte superior, de forma similar a un fotodiodo, se hace en forma de rejilla con bandas distribuidas adecuadamente sobre toda la superficie con objeto de disminuir la resistencia de contacto, y por tanto la resistencia serie R_s , tratando a la vez de no restar área expuesta a la radiación.

5.5.4 Diodos emisores de luz

Los diodos emisores de luz o LED³ tienen su fundamento en el proceso inverso a los diodos fotodetectores. Así, cuando un diodo se polariza en directo los portadores minorita-

³ Nota: El acrónimo LED procede del nombre en lengua inglesa "Light Emitting Diode".

rios, electrones y huecos, una vez inyectados en las regiones neutras de signo opuesto, acaban finalmente recombinándose con los portadores mayoritarios de la región. En estos procesos de recombinación un electrón pasa siempre desde la banda de conducción a la de valencia por lo que en el caso ideal debería existir emisión de un fotón de energía $h\nu$ igual a la de la banda prohibida (fig. 5.24). En la práctica, sin embargo, las transiciones radiativas están también acompañadas de transiciones no radiativas en las cuales la energía de los portadores inyectados acaba transformándose en un aumento del estado de vibración de la red (calor).

La estructura de las bandas de energía del semiconductor es un factor determinante del tipo de transición, ya que según sea esta estructura las transiciones pueden ser directas o indirectas. En las *transiciones directas* el electrón pasa a la banda de valencia sin cambio de momento. Los semiconductores que presentan este tipo de transición se denominan de "gap" directo, según vimos en el capítulo primero (sec. 1.4.4). Por contra, en las *transiciones indirectas* el electrón cambia su momento en la transición mediante la participación de un *fonón* que lógicamente cambia el estado de vibración de la red. Estas transiciones se presentan en los semiconductores de "gap" indirecto. Ambas transiciones son radiativas pero la probabilidad de las indirectas es mucho menor que la probabilidad de las directas. Por tanto otros procesos competitivos no radiativos son mucho más frecuentes en los semiconductores de "gap" indirecto que en los de "gap" directo.

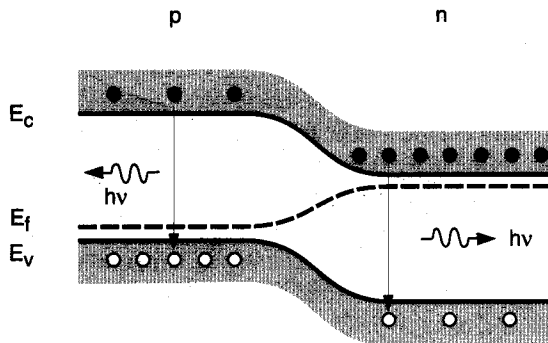


Fig. 5.24. Procesos de recombinación de minoritarios en una unión p-n polarizada en directo, con emisión de fotones (diodo LED).

Entre los semiconductores más conocidos, el silicio y el germanio son de "gap" indirecto, mientras que el arseniuro de galio y otros compuestos de los grupos III y V (compuestos III-V) presentan "gap" directo. El GaAs tiene una banda prohibida de 1.43 eV y por ello se

usa como material apropiado para LED en la región de infrarrojo próximo ($\approx 0.9 \mu\text{m}$). Esta región del espectro está siendo muy utilizada actualmente en las comunicaciones ópticas mediante fibra óptica, por lo que los LED de GaAs están alcanzando un gran desarrollo.

Existen otros semiconductores compuestos, como el GaP o el AlAs, que tienen una banda prohibida con energía más elevada, aunque son de "gap" indirecto. Sin embargo, se puede construir LED's que emitan en el visible partiendo de estos materiales si se consigue eliminar las transiciones no radiativas. Estos materiales también son utilizados para formar compuestos ternarios con el GaAs, del tipo $\text{GaAs}_{1-x}\text{P}_x$ con una fracción, x , pequeña. Con ello se consigue aumentar notablemente la anchura de la banda prohibida del GaAs (hasta un valor de 2 eV, aproximadamente) manteniéndose al mismo tiempo el tipo de transiciones directas.

5.5.5. Diodos láser

Cuando la densidad de corriente en un diodo emisor de luz aumenta hasta un cierto límite, se puede presentar la emisión de radiación en forma de *láser*, es decir, luz monocromática, coherente y confinada en un haz no divergente. En la mayoría de los LED generalmente existe una densidad de corriente umbral por encima de la cual aparece este efecto de emisión estimulada. La condición necesaria para que se presente emisión estimulada es que se produzca una alta inversión de población de portadores en las regiones neutras, como consecuencia de la inyección de minoritarios desde el lado opuesto. Al mismo tiempo ha de existir también una elevada densidad de fotones, los cuales pueden proceder del propio diodo a través de la recombinación de minoritarios en las regiones neutras. Esto último se consigue confinando la región de recombinación a zonas muy estrechas, haciendo que la luz emitida sufra múltiples reflexiones internas antes de salir del diodo (*cavidad resonante de Fabry-Perot*). Un buen confinamiento de la luz se puede obtener por ejemplo si existe una diferencia muy acusada entre el índice de refracción de la región de recombinación (generalmente esta región es la de agotamiento) y el índice de las zonas neutras.

La fig. 5.25 muestra un esquema de bandas de energía del proceso de emisión láser en una unión p-n polarizada en directo, con un voltaje suficientemente elevado de forma que la corriente esté por encima del valor umbral. Una corriente elevada se puede conseguir utilizando semiconductores degenerados de forma que el nivel de Fermi a cada lado de la unión esté por encima (debajo) del borde de la banda de conducción (valencia). En estas condiciones se obtiene una unión abrupta con un voltaje de contacto elevado (fig. 5.25a). Si se polariza en directo con un voltaje V la unión funciona como un diodo normal, inyectando electrones en el lado p y huecos en el lado n (fig. 5.25b). Cuando el voltaje es suficientemente elevado la concentración de minoritarios a ambos lados de la unión es tan elevada que en las proximidades de la unión (región de espesor x en la fig. 5.25c) se dan condiciones de inversión de población, es decir, la concentración de electrones en la banda de conducción en el lado p es mucho más elevada que la correspondiente al equilibrio térmico. Obviamente, para los huecos

se tiene una situación análoga. Ocurre entonces la emisión láser por recombinación de portadores minoritarios en esa región (fig. 5.25c).

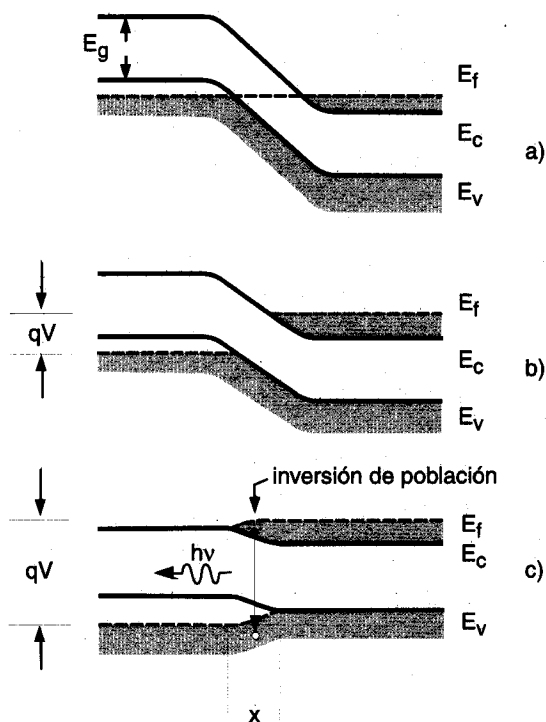


Fig. 5.25. Diagrama de energía de un proceso de emisión láser: a) Unión p-n en equilibrio formada por dos semiconductores degenerados, b) Unión p-n polarizada en directo, y c) Polarización en condiciones de alta inyección de mayoritarios, con emisión estimulada de luz (láser) en la región de inversión de poblaciones (de anchura x).

La emisión láser se ha observado en semiconductores principalmente de "gap" directo, lo cual es explicable si se tiene en cuenta que en ellos las transiciones radiativas son las más probables. Entre estos semiconductores se encuentra en lugar prominente el GaAs y otros compuestos ternarios y cuaternarios de los grupos III y V, como el $\text{Al}_x\text{Ga}_{1-x}\text{As}$, el $\text{Al}_x\text{Ga}_{1-x}\text{As}_y\text{Sb}_{1-y}$, etc. los cuales tienen un valor de E_g creciente a medida que aumentan las fracciones x ó y de la composición. Todos estos materiales son ampliamente utilizados en la preparación de diodos láser. Debido a la posibilidad de modular la intensidad luminosa me-

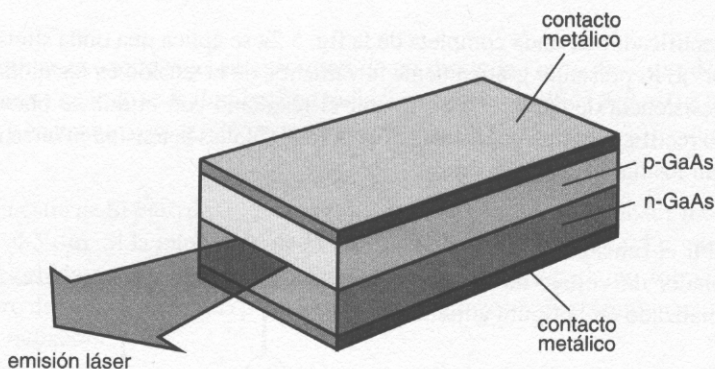


Fig. 5.26. Estructura de un diodo láser de GaAs, tipo homounión.

diente el voltaje aplicado, las aplicaciones más inmediatas de estos diodos se encuentra en el campo de las comunicaciones por fibra óptica y en el procesado de señales luminosas mediante óptica integrada.

En la fig. 5.26 se muestra una representación esquemática de la estructura de un diodo láser del tipo *homounión* formado por dos semiconductores de GaAs de tipos p y n. Las caras frontales son perfectamente paralelas y perpendiculares a la unión con objeto de obtener una cavidad tipo Fabry-Perot. Este tipo de láser fue el primero de los desarrollados hacia los años 60. Posteriormente se fabricaron nuevas estructuras formadas por uniones dobles del GaAs con compuestos ternarios y también por heterouniones (unión de semiconductores con diferentes composiciones) que dan emisión láser con una corriente umbral mucho más baja.

CUESTIONES Y PROBLEMAS

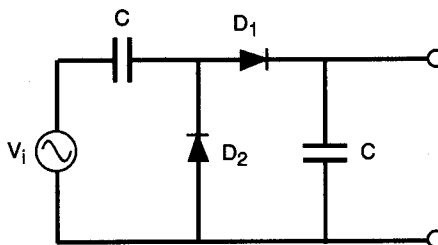
- 5.1 Determinar el punto de funcionamiento de un diodo polarizado en directo a través de una resistencia de carga de 9 ohmios mediante una fuente de alimentación de 1 V con resistencia interna de 1 ohmio (corriente de saturación del diodo, $I_0 = 10^{-12}$ A). Si la

corriente máxima que puede soportar el diodo es de 75 mA. ¿Cuál es la resistencia mínima necesaria para limitar la corriente a este valor?

- 5.2 En el rectificador de onda completa de la fig. 5.2a se aplica una onda sinusoidal v_i a la entrada. a) Representar gráficamente la variación de la tensión en los diodos D_1 y D_2 y en la resistencia de carga. b) Comparar el resultado con el que se obtendría con el circuito rectificador tipo puente de la fig. 5.3a. ¿Cuál es la tensión inversa máxima que soportan los diodos en cada caso?

- 5.3 Describir el funcionamiento del circuito doblador de voltaje de media onda esquematizado en la figura adjunta.

- 5.4 Describir el efecto de filtrado que produce un condensador de capacidad elevada colocado a la salida de un puente rectificador de onda completa.



- 5.5 Dos diodos se conectan en oposición directamente a la salida de una fuente de alimentación de 6 V. a) Calcular la corriente y la tensión que soporta cada diodo. b) Suponiendo que la tensión de ruptura de los diodos es de 5.8 V, ¿cuál sería la corriente en el circuito?. Utilícese para la corriente de saturación de los diodos el valor de $I_0 = 1 \mu\text{A}$.

- 5.6 A menudo se representa el comportamiento de un diodo Zener por un circuito equivalente formado por un generador de tensión V_z (V_z es la tensión de ruptura del diodo) y una resistencia serie, R_z (R_z es la resistencia dinámica), ambos de valor constante. De acuerdo con esta aproximación, calcular: a) La variación de la tensión de salida V_o con la intensidad I en el circuito de carga de una fuente de alimentación de tensión V_i estabilizada con el diodo Zener. b) La fuente de alimentación equivalente obtenida mediante el teorema de Thévenin.

- 5.7 Un diodo Zener de 60 V de voltaje de ruptura funciona entre 2 y 40 mA para estabilizar una fuente de alimentación de 100 V y resistencia interna de 10 ohmios. Calcular la resistencia serie necesaria para mantener una salida constante en la carga. ¿Cuál es la corriente máxima que puede suministrar la fuente?

- 5.8 Si a la fuente del problema anterior se le conecta una resistencia de carga de 6000 ohmios, ¿entre qué límites puede variar la tensión del generador manteniéndose constante la tensión de salida?

- 5.9** Un diodo túnel de Si está dopado de forma que $N_a = N_d$, siendo la concentración de impurezas de 1 por cada 10^3 átomos de Si. Calcular la altura de la barrera de la unión y la anchura de la región de carga espacial.
- 5.10** Determinar el coeficiente de absorción de un pieza de GaAs de $7 \mu\text{m}$ de espesor, sabiendo que cuando se la ilumina con radiación monocromática de $\lambda = 0.8 \mu\text{m}$ refleja un tercio y transmite otro tercio de la radiación.
- 5.11** Una muestra de Si puro de $0.1 \mu\text{m}$ de espesor se ilumina con luz con luz monocromática de $\lambda = 0.5 \mu\text{m}$. Si la intensidad de la luz es de 50 mW , calcular: a) La intensidad de la luz absorbida en el semiconductor. b) La intensidad disipada en forma de calor. c) Número de fotones emitidos por segundo en los procesos de recombinación originados por la radiación.
- 5.12** Un fotodiodo se ilumina con un haz puntual a una distancia x de la unión. Calcular la variación de la corriente originada en el diodo con la distancia x .
- 5.13** A partir de la curva característica I-V de una célula solar dada por la ec. [5.15], a) Determinar gráficamente el punto de operación cuando se conecta una resistencia R_L entre los terminales de salida de la célula. b) Determinar el valor que ha de tener R_L para conseguir la máxima potencia de salida.
- 5.14** A partir del circuito equivalente de una célula solar mostrado en la fig. 5.22, discutir el efecto de la resistencia R_s en el circuito de salida.

CAPITULO VI

TRANSISTORES BIPOLARES

Uno de los dispositivos de estado sólido que mayor impacto ha causado en el desarrollo de los circuitos electrónicos es quizás el transistor. Descubierto en 1947 por un grupo de científicos de la compañía Bell Telephone de EE.UU., pronto alcanzó una enorme popularidad sustituyendo a los antiguos elementos de vacío debido a su menor tamaño y consumo de potencia. Al mismo tiempo se abrieron nuevas posibilidades de aplicación en circuitos en los cuales los transistores ejecutan funciones cada vez más complejas. Actualmente existe una familia muy amplia de dispositivos transistores, cada uno de ellos con características muy específicas. En este capítulo nos referiremos a los *bipolares*, denominados así porque en los procesos de conducción que ocurren durante el funcionamiento de estos dispositivos participan portadores de ambos signos, es decir, huecos y electrones. En capítulos posteriores se estudiará el comportamiento de otros tipos, en particular los denominados de efecto de campo, de carácter unipolar.

6.1. TRANSISTORES BIPOLARES DE UNION: DESCRIPCION Y NOMENCLATURA

El transistor bipolar de unión, muy a menudo denominado simplemente *transistor*, está constituido por la unión de tres semiconductores, de carácter p y n alternativamente. El semiconductor del centro suele ser muy estrecho y se denomina *base*. Los otros dos semiconductores, de signo opuesto a la base se denominan, de acuerdo con las funciones que ejecutan, *emisor* y *colector*. Cada uno de estos semiconductores lleva un contacto metálico con un hilo de salida al exterior. Se trata pues de un dispositivo de tres terminales.

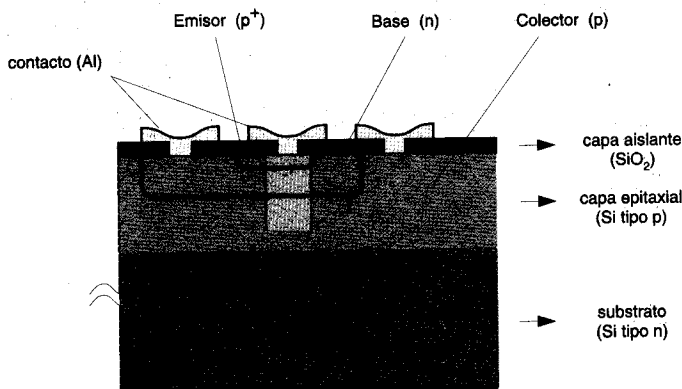


Fig.6.1. Estructura de un transistor bipolar tipo pnp de silicio, mostrando la disposición de cada una de las regiones y de los contactos hacia el exterior. Se indica también la zona donde tiene lugar la acción del transistor (área sombreada).

En circuitos integrados, la preparación de transistores y otros componentes del circuito se lleva a cabo mediante la *tecnología planar*, la cual se describe en el último capítulo de este tratado. Según esta tecnología, las regiones p y n del transistor se obtienen por difusión de las impurezas correspondientes sobre la superficie de una pieza única del material semiconductor cortado en forma de oblea. Sobre cada una de las regiones se deposita después el contacto metálico, de tipo óhmico, para conectar los terminales de salida. La disposición resultante para cada una de estas regiones, así como los contactos metálicos hacia el exterior, viene indicada en la figura 6.1. La figura muestra la sección transversal de una oblea de silicio, la cual lleva sobre su superficie una capa del mismo material (capa epitaxial) depositada por procedimientos químicos. Es esta capa donde se distribuyen las diferentes regiones del transistor. Los contactos metálicos para hacer conexión con los hilos de salida se depositan en forma de película delgada sobre la superficie del semiconductor. En la figura se muestra en tono claro la zona activa en la cual ocurre la acción del transistor.

La fig. 6.2 ilustra la disposición de los semiconductores en los dos tipos posibles de transistores, pnp y npn, así como el símbolo correspondiente utilizado en los esquemas de circuitos. Se incluye también las polaridades de las tensiones y las direcciones de las corrientes en cada terminal cuando el transistor está operando en su modo normal o, dicho de otra forma, en la región *activa*. En este modo de operación la unión entre emisor y base (unión de emisor) está polarizada en directo, mientras que la unión entre base y colector (unión de colec-

tor) se polariza en inverso. En un transistor pnp, esto quiere decir que el emisor debe estar polarizado positivamente respecto de la base, y a su vez la base debe ser más positiva que el colector. Para determinar el signo de las corrientes en el esquema de la fig. 6.2 conviene seguir la convención más simple que considera que todas las corrientes son positivas cuando el transistor está polarizado en la región activa. A este respecto conviene señalar que, en lo que se refiere a las tensiones aplicadas, la segunda letra del subíndice se refiere al terminal de referencia. Así, por ejemplo, $V_{EB} > 0$ es equivalente a $V_E - V_B > 0$ y quiere decir que el emisor es positivo respecto a la base. Nótese que los signos de las corrientes y de las tensiones de un transistor pnp son opuestos a los correspondientes en un transistor npn cuando ambos operan en la región activa. Por este motivo en este capítulo nos centraremos sobre todo en los primeros, aunque todas las conclusiones serán válidas para ambos tipos, con la salvedad de hacer el consiguiente cambio de signo, tanto en las corrientes y tensiones como en el de los portadores.

Según sea el signo de las tensiones aplicadas en las uniones de emisor y colector, se dice que el transistor tiene cuatro regiones de funcionamiento, indicadas en el esquema de la fig. 6.3 para un transistor pnp. Como veremos más adelante, la *región activa*, ya mencionada (con $V_{EB} > 0$ y $V_{BC} > 0$), es la más común y en ella el factor de amplificación es el más elevado. Por contra en la *región inversa* ($V_{EB} < 0$ y $V_{BC} < 0$) las polaridades están invertidas con respec-

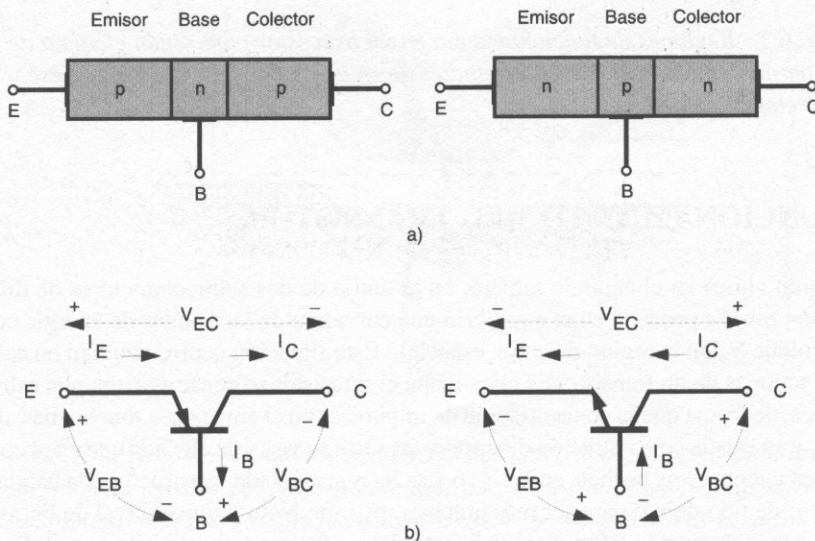


Fig. 6.2. Estructura (a) y símbolo de circuitos (b) de transistores pnp y npn. Cuando las tensiones aplicadas tienen la polaridad indicada en la figura el transistor funciona en la región activa. Las corrientes en los terminales tienen entonces el sentido indicado que, por convención, se toma como positivo.

to a la región activa, de forma que el colector funciona como emisor y éste como colector. Debido a las características de diseño, el transistor opera en estas condiciones con una ganancia mucho más baja. En las otras dos regiones, *saturación* ($V_{EB} > 0$ y $V_{BC} < 0$) y *corte* ($V_{EB} < 0$ y $V_{BC} > 0$), el transistor funciona prácticamente como una resistencia entre los terminales de emisor y colector, con un valor bajo o elevado según sea el caso.

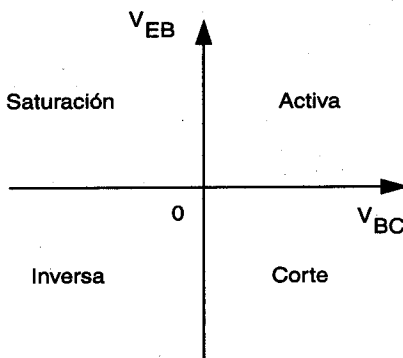


Fig. 6.3. Regiones de funcionamiento en un transistor pnp, según el signo de la tensión aplicada entre los terminales de emisor y base, V_{EB} , y los de base y colector, V_{BC} .

6.2. FUNCIONAMIENTO DEL TRANSISTOR

Según vimos en el capítulo tercero, en la unión de dos semiconductores de diferente signo (unión p-n) se produce en el equilibrio una curvatura de las bandas de energía con una caída de voltaje V_o en la región de carga espacial. Esta situación ocurre también en cada una de las dos uniones de un transistor. Típicamente el transistor se construye con una estructura no simétrica, de forma que la concentración de impurezas en el emisor sea mucho más alta que en la base, y en ésta la concentración de impurezas sea a su vez más elevada que en el colector. Teniendo en cuenta estos hechos, en la fig. 6.4 se ha representado la estructura de bandas para un transistor de tipo pnp, o para ser más precisos, p^+-n-p . Nótese que el nivel de Fermi en el equilibrio es constante a lo largo de toda la estructura. Además, de acuerdo con el diseño, la curvatura de las bandas es más abrupta, y la región de carga espacial más estrecha en la unión de emisor que en la unión de colector. En la fig. 6.4 se ha incluido también las curvas de distribución en energía de electrones y huecos, $n(E)$ y $p(E)$, en cada una de las regiones una vez que se establece el equilibrio termodinámico. Según se indica, existe una alta densidad de huecos en el emisor y colector y de electrones en la base.

6.2.1. Operación en la región activa

Cuando el transistor pnp se polariza en la región activa, esto es $V_{EB} > 0$ y $V_{BC} > 0$, la unión de emisor queda polarizada en directo mientras que la unión de colector está polarizada en inverso. Recuérdese que en polarización directa el voltaje V_{EB} debe ser de unas décimas de voltio, esto es, menor que el potencial de contacto asociado a la unión, mientras que en polarización inversa el potencial V_{BC} puede oscilar en varias decenas de voltio. Desde un punto de vista energético la aplicación del voltaje V_{EB} significa que la barrera emisor/base disminuye en una pequeña cantidad, dada por qV_{EB} , y la barrera base/colector aumenta en mayor proporción, en la cantidad qV_{BC} , según se muestra en fig. 6.5a. En consecuencia, si se consideraran las uniones separadamente debería existir una corriente elevada dominada por los portadores mayoritarios en la unión de emisor, y simultáneamente una corriente relativamente pequeña y dominada por los portadores minoritarios en la unión de colector. Sin embargo, la proximidad de las uniones de emisor y de colector hace que la situación en el transistor sea muy diferente.

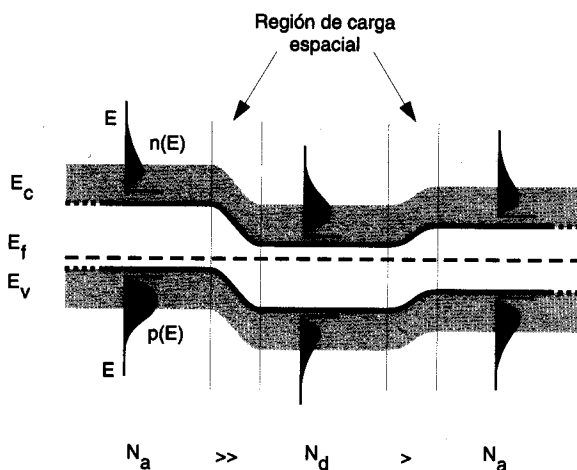


Fig.6.4. Estructura de bandas de un transistor típico tipo p^+-n-p en equilibrio, sin ninguna tensión de polarización aplicada en los terminales. Las curvas sombreadas muestran la función de distribución de los electrones y huecos en cada una de las regiones del transistor.

Si el emisor está mucho más dopado que la base (unión p^+-n) la corriente a través de la unión de emisor está formada fundamentalmente por los huecos del emisor que se difunden hacia la base. A esta corriente de huecos la denominaremos $I_{E,p}$. Un aspecto notable del comportamiento del transistor está relacionado con esta corriente, ya que si la anchura de la

base es suficientemente estrecha, tal como se ha señalado anteriormente (de hecho, la anchura ha de ser menor que la longitud de difusión de los huecos en la base), **la mayor parte de estos huecos alcanza la unión de colector antes de recombinarse con los electrones de la base**. Hay que hacer notar que la fracción de huecos que cruza la base sin llegar a recombinarse con los electrones de esta región está sometida al potencial negativo que existe entre base y colector, por lo que esta fracción de huecos termina finalmente en el colector¹.

Junto a esta corriente de huecos que atraviesa la base desde el emisor, en la unión de emisor hay que considerar también el desplazamiento de los electrones que circula en sentido opuesto al de los huecos. Una fracción de la corriente de electrones que circula por la base se recombina con los huecos en la propia base, según acabamos de ver, y el resto pasa al emisor. Si la unión de emisor es del tipo $p^+ - n$, el flujo de electrones que atraviesa la unión de emisor en la dirección del emisor es mucho menor que el de huecos que circula en sentido opuesto. Esta corriente de electrones será denominada $I_{E,n}$, de forma que la corriente total de emisor, I_E , será:

$$I_E = I_{E,p} + I_{E,n} \quad [6.1]$$

En la unión de colector (polarizada en inverso), aparte de la corriente principal de huecos que procede del emisor, existe también sendas corrientes de electrones y huecos minoritarios que viajan en sentidos opuestos. Si la base está mucho más dopada que el colector la corriente de minoritarios en la unión de colector está formada fundamentalmente por electrones que parten desde el colector. Por esta razón, a menudo esta corriente se designa por $I_{C,n}$, implicando con ello que la corriente de huecos minoritarios es muy pequeña. Si además $I_{C,p}$ representa la corriente principal de huecos que procede del emisor y atraviesa la base, la corriente total que circula por el colector, I_C , vendrá dada aproximadamente por:

$$I_C = I_{C,p} + I_{C,n} \quad [6.2]$$

Quizás, el efecto más importante del transistor es la capacidad para inyectar en el colector una corriente elevada procedente del emisor a través de la base, a pesar de que la unión de colector está polarizada en inversa. En una estructura $p^+ - n - p$ esta corriente está formada por huecos, fundamentalmente. Esto sólo se consigue si la base es suficientemente estrecha de forma que la corriente inyectada procedente del emisor alcance el colector. Si el emisor estuviese distanciado del colector la corriente de emisor se recombinaría en la base y el dispositivo sería equivalente a una simple unión de dos diodos conectados en oposición. El efecto anterior es además el que da lugar a la terminología "emisor" y "colector", para nombrar a la región que realmente emite y recoge, respectivamente, la corriente principal en el transistor.

¹ **Nota:** Resulta de gran ayuda para conocer el sentido de movimiento de los electrones y de los huecos frente a una barrera de potencial la imagen descrita en el capítulo primero que considera a los electrones como partículas pesadas moviéndose hacia abajo en la banda de conducción y los huecos como burbujas de un fluido que tienden a flotar hacia arriba en la banda de valencia.

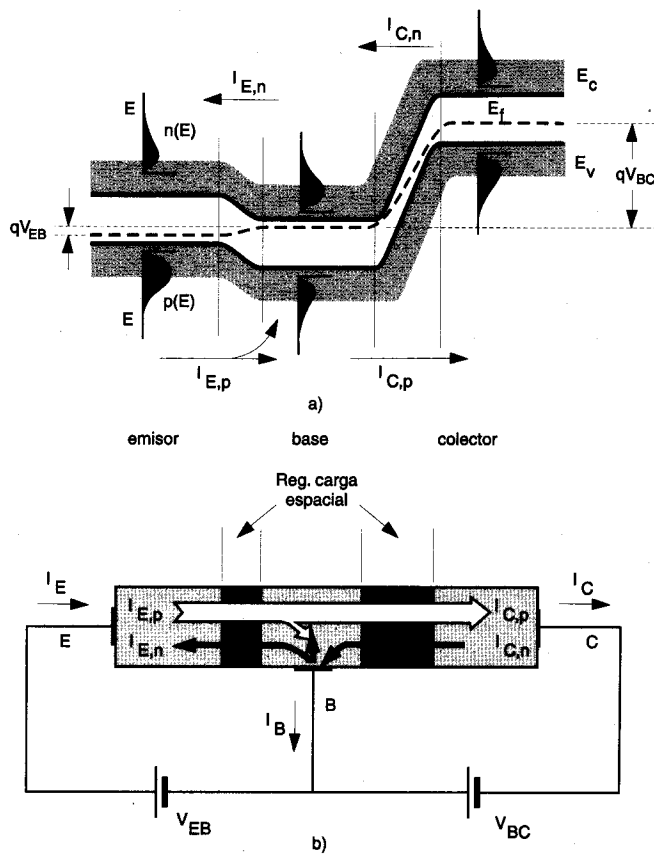


Fig. 6.5. a) Esquema de la estructura de bandas de un transistor pnp polarizado en la región activa, mostrando las curvas de distribución en energía de electrones y huecos. b) Esquema del movimiento de huecos (en color claro) y de electrones (en color oscuro) a través del transistor cuando está polarizado en la región activa.

En la fig. 6.5b se presenta un esquema de las diferentes contribuciones a las corrientes que circulan por el interior de un transistor tipo pnp polarizado en la región activa. En trazo claro se ha señalado la corriente principal de huecos desde el emisor al colector pasando por la base. La contribución de los electrones, cuyo movimiento se realiza en sentido opuesto al de los huecos, se ha representado en trazo oscuro.

6.2.2. Parámetros de diseño del transistor.

Las ecuaciones [6.1] y [6.2] permiten definir algunos parámetros útiles para determinar el comportamiento del transistor cuando funciona en la región activa. Así, por ejemplo, se define *el factor de transporte de base*, α_T , como la relación entre la corriente de huecos que se difunde a través de la unión de colector y la corriente de huecos inyectada a través de la unión de emisor, esto es:

$$\alpha_T = \frac{I_{C,p}}{I_{E,p}} \quad [6.3]$$

En el caso ideal, en el que no existiese recombinación de huecos en la base, α_T sería prácticamente la unidad. De hecho α_T es siempre algo menor que la unidad, y su valor está determinado fundamentalmente por el cociente entre la anchura de la base w y la longitud de difusión de los huecos, L_h , de acuerdo con la ecuación aproximada:

$$\alpha_T = 1 - \frac{w^2}{2L_h^2} \quad [6.4]$$

Con objeto de aumentar α_T , el cociente w/L_h debe ser lo más bajo posible. Por esta razón los transistores se fabrican con una base muy estrecha, con una longitud incluso por debajo de la micra, es decir, menor que la longitud de difusión de los huecos.

Asimismo, se define también la *eficiencia del emisor*, γ , como la proporción de la corriente de huecos inyectada en la unión de emisor en relación a la corriente total de emisor:

$$\gamma = \frac{I_{E,p}}{I_E} = \frac{I_{E,p}}{I_{E,p} + I_{E,n}} \quad [6.5]$$

Análogamente, bajo ciertas aproximaciones, se cumple:

$$\gamma = \left[1 + \frac{D_e}{D_h} \frac{N_B}{N_E} \frac{w}{L_h} \right]^{-1} \quad [6.6]$$

siendo N_E y N_B las concentraciones de dopantes (aceptores y donadores) en las regiones de emisor y base, respectivamente. Evidentemente, γ mide la capacidad del emisor para inyectar huecos en la base. Como veremos más adelante, interesa que γ sea próximo a la unidad, ya que entonces el factor de ganancia es más elevado. En una estructura pnp esto se consigue dopando fuertemente la región de emisor con impurezas aceptoras (estructura $p^+ - n - p$) de forma que la corriente $I_{E,p}$ sea mucho más elevada que la corriente $I_{E,n}$ en la unión de emisor (véase la fig. 6.5b).

6.2.3. Parámetros de funcionamiento como amplificador

Debido a que el transistor es un dispositivo de tres terminales, su utilización en circuitos amplificadores de señales alternas exige que uno de los terminales sea común a los circuitos de entrada y salida de señal. De las tres *configuraciones* posibles, las más empleadas son la que tiene el terminal *de base común* a los circuitos de entrada y salida (fig. 6.6a) y la que tiene el terminal *de emisor común* a ambos circuitos (fig. 6.6b). Atendiendo a cada una de estas dos configuraciones se define el correspondiente *factor de ganancia en corriente*, parámetro que resulta muy útil para conocer el comportamiento del transistor cuanto trabaja en la **región activa**. Así para la configuración de base común se utiliza el factor α_{dc} definido como²:

$$\alpha_{dc} = \frac{I_{C.p}}{I_E} \quad [6.7]$$

es decir, el cociente entre la corriente de huecos que alcanza el colector y la corriente total de emisor. Dado que en la región activa $I_{C.p} \approx I_C$, resulta $\alpha_{dc} \approx I_C/I_E$. Por esta razón, a menudo también se define un nuevo factor a partir del cociente incremental:

$$\alpha = \left. \frac{\Delta I_C}{\Delta I_E} \right|_{V_{BC}} \quad [6.8]$$

útil cuando se trabaja en circuitos de amplificación con pequeñas señales de corriente, ΔI_C , superpuestas a un valor continuo, I_C (el correspondiente a la tensión V_{BC}).

Análogamente, en la configuración de emisor común se utiliza el factor de ganancia β_{dc} definido a través de la relación:

$$\beta_{dc} = \frac{I_C}{I_B} \quad [6.9]$$

Del mismo modo, para señales pequeñas de corriente se define también el factor:

$$\beta = \left. \frac{\Delta I_C}{\Delta I_B} \right|_{V_{EC}} \quad [6.10]$$

es decir el cociente incremental entre la corriente de colector y la corriente de base para una tensión V_{EC} dada. Todos estos parámetros, que miden la relación entre una corriente en el terminal activo del circuito de salida (colector) y la correspondiente al terminal activo del

² Nota: El subíndice dc se emplea para significar que se trata de parámetros en corriente continua.

circuito de entrada (emisor o base, según el caso), son de gran utilidad en la descripción del comportamiento del transistor en la región activa, ya que fuera de esta región las relaciones entre las corrientes no son lineales. Por esta razón los valores α_{dc} y β_{dc} suelen venir referidos a la región activa, aunque es muy conveniente en cualquier caso definir el punto de operación del transistor.

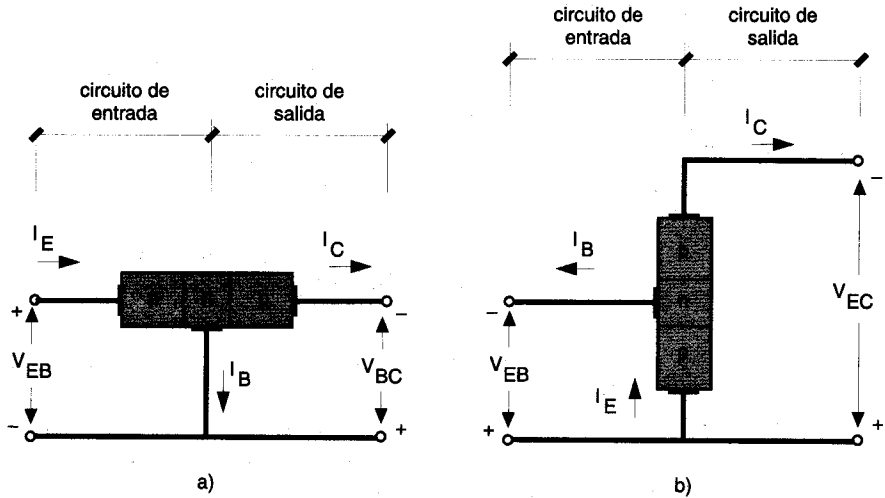


Fig. 6.6. a) Configuración del transistor en base común polarizado en la región activa. b) Configuración de emisor común, también polarizado en la región activa.

Dado que los parámetros α_T y γ dependen fundamentalmente de las características de diseño del propio transistor, es lógico esperar que los valores α_{dc} y β_{dc} dependan de estos parámetros. Así la relación [6.7] permite obtener una ecuación para calcular el valor de α_{dc} . En efecto:

$$\alpha_{dc} = \frac{I_{C,p}}{I_{E,p} + I_{E,n}} = \frac{I_{C,p}}{I_{E,p}} \frac{I_{E,p}}{I_{E,p} + I_{E,n}} = \alpha_T \gamma \quad [6.11]$$

Similarmemente, a partir de la relación [6.9] y utilizando la ley de Kirchhoff para el nudo de corriente formado en el transistor:

$$I_E = I_B + I_C \quad [6.12]$$

se obtiene:

$$\beta_{dc} = \frac{I_C}{I_B} = \frac{I_C}{I_E - I_C} = \frac{I_C / I_E}{1 - (I_C / I_E)} \approx \frac{\alpha_{dc}}{1 - \alpha_{dc}} \quad [6.13]$$

la cual relaciona entre sí los dos factores de ganancia en corriente continua. Utilizando ecuaciones similares se obtiene análogamente para los parámetros de pequeña señal:

$$\beta = \frac{\alpha}{1 - \alpha} \quad [6.14]$$

A partir de estos parámetros es posible relacionar directamente las corrientes en los circuitos de entrada y salida en las dos configuraciones señaladas en la fig. 6.6. Así, para la configuración de base común se obtiene de las ecs. [6.2] y [6.7]:

$$I_C = I_{C,p} + I_{C,n} = \alpha_{dc} I_E + I_{C,n} \quad [6.15]$$

En esta ecuación, $I_{C,n}$ se interpreta como la corriente total en la unión de colector (coincide aproximadamente con la corriente de minoritarios) cuando el emisor está en circuito abierto ($I_E = 0$). Frecuentemente el valor de $I_{C,n}$ se representa por I_{CBO} , significando con los dos primeros subíndices (CB) los dos terminales a través de los cuales fluye la corriente y con el tercero (O) el estado de circuito abierto del tercer terminal. De acuerdo con esta notación la corriente de colector suele escribirse como:

$$I_C = \alpha_{dc} I_E + I_{CBO} \quad [6.16]$$

Así mismo, para la configuración de emisor común tendremos:

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

Resolviendo para I_C :

$$I_C = \frac{\alpha_{dc}}{1 - \alpha_{dc}} I_B + \frac{I_{CBO}}{1 - \alpha_{dc}}$$

o bien, utilizando la ecuación aproximada [6.13]

$$I_C \approx \beta_{dc} I_B + I_{CEO} \quad [6.17]$$

siendo I_{CEO} la corriente de emisor a colector con la base en circuito abierto ($I_B = 0$), dada por:

$$I_{CEO} = \frac{I_{CBO}}{1 - \alpha_{dc}} \quad [6.18]$$

Dado que en la región activa el parámetro α_{dc} es un factor muy próximo a la unidad, β_{dc} es un factor mucho mayor que 1. Así por ejemplo, si $\alpha_{dc} = 0.99$, β_{dc} vale 99. Generalmente el valor de β_{dc} de un transistor oscila entre 60 y 300. Por tanto, según la ec. [6.18], I_{CEO} es mucho mayor que I_{CBO} . Este hecho es explicable si se tiene en cuenta que, incluso cuando la base está en circuito abierto, la tensión aplicada entre emisor y colector polariza la unión de emisor débilmente en directo, lo cual hace que haya una cierta corriente de huecos del emisor hacia el colector a través de la base. Hay que resaltar que de acuerdo con [6.17] una pequeña variación de la corriente de base determina una variación elevada de la corriente de colector cuando el transistor funciona en la región activa. Quizás es éste uno de los aspectos más interesantes del transistor en su aplicación en circuitos amplificadores, según se estudia en el capítulo 9.

6.3. CURVAS CARACTERISTICAS I-V DE LOS TRANSISTORES.

El cálculo teórico de la relación intensidad-voltaje en cada una de las dos uniones que componen el transistor es relativamente complejo debido a la interrelación entre las corrientes en una y otra unión. Sin embargo, el cálculo es en esencia muy similar al que se siguió en el capítulo tercero para determinar la característica I-V en un diodo. Este cálculo incluye la determinación del exceso de portadores minoritarios que ha pasado a cada lado de la unión como consecuencia de la aplicación de un voltaje de polarización. Según vimos en el capítulo tercero, la corriente a través de una unión p-n resulta ser proporcional al gradiente de portadores minoritarios que existe a cada lado de la unión, mostrando una dependencia exponencial con el voltaje aplicado (véase apartado 3.3.3).

6.3.1. Modelo de Ebers-Moll (*)

En el caso del transistor, se puede demostrar que las corrientes que pasan a través de la unión de emisor y de colector tienen una dependencia exponencial con el voltaje aplicado a ambos lados de cada una de las uniones (V_{EB} y V_{BC} , respectivamente). Así, por ejemplo, es posible generalizar la ec. [6.16], en principio sólo válida para la región activa, para aquellos casos en que la unión de colector esté polarizada a una tensión cualquiera, positiva o negativa. Escribamos primero la ec. [6.16] en la forma:

$$I_C = \alpha_F I_E + I_{CO}$$

donde el subíndice "F" del coeficiente α expresa que este parámetro se refiere ahora al sentido directo ("forward") de la corriente principal de huecos de emisor a colector. En la igualdad anterior se ha hecho además $I_{CBO} = I_{CO}$. La utilidad de este cambio de nomenclatura se hará evidente más adelante. Si queremos que esta ecuación incluya también el caso en que la unión de colector esté polarizada en directo, debemos poner:

$$I_C = \alpha_F I_E - I_{CO} [\exp (qV_{CB} / kT) - 1] \quad [6.19]$$

El segundo término de esta igualdad pone de manifiesto la dependencia exponencial de la corriente a través de la unión de colector de acuerdo con la ecuación del diodo. Obsérvese que esta ecuación incluye a la anterior como caso particular, ya que en la región activa el término exponencial se reduce prácticamente a cero.

Podemos plantear una ecuación similar a la ec. [6.19] para el caso extremo del transistor funcionando en la región inversa ($V_{EB} < 0$ y $V_{BC} < 0$). En este caso existiría también una corriente de huecos inyectada desde el colector al emisor, acompañada en el emisor de la corriente de minoritarios en la unión de emisor que se encuentra ahora polarizada en inversa. Si la ecuación se extiende al caso de polarización positiva para la unión de emisor (es decir $V_{EB} > 0$), la corriente de emisor vendrá dada por:

$$I_E = \alpha_R I_C + I_{EO} [\exp (qV_{EB} / kT) - 1] \quad [6.20]$$

En esta ecuación, el coeficiente α_R es el *factor de ganancia* en corriente, ec. [6.7], cuando la corriente principal de huecos tiene sentido opuesto ("reverse") al caso anterior, es decir de colector a emisor. El segundo término muestra también la dependencia exponencial de la corriente con el voltaje en la unión de emisor. El factor I_{EO} tiene un significado similar a la corriente I_{CO} , esto es, representa la corriente de saturación en la unión de emisor cuando el terminal en el lado opuesto a la unión (el colector en este caso) está en circuito abierto. En un transistor tipo p^+-n-p , tanto α_R como I_{EO} son mucho menores que las correspondientes magnitudes en sentido directo.

En conjunto, las dos ecuaciones anteriores abarcan todas la situaciones posibles que se pueden presentar en el funcionamiento del transistor. Su importancia fue pronto reconocida por Ebers y Moll, de ahí que el modelo de circuito equivalente para el transistor, dado en la fig. 6.7a, lleve su nombre. El circuito incluye dos diodos en oposición, que dan cuenta de los términos exponenciales de las dos ecuaciones anteriores. Estos diodos tienen conectado en paralelo sendas fuentes de corriente, de valor $\alpha_F I_E$ y $\alpha_R I_C$, que representan específicamente la "acción" del transistor descrita más arriba, esto es, la inyección de huecos de emisor a colector (sentido directo) o de colector a emisor (sentido inverso), según el caso.

A veces, interesa expresar el valor de las fuentes de corriente del circuito de la fig. 6.7a en función de la tensión aplicada a cada una de las dos uniones que componen el transistor. Así, si en la ec. [6.19] sustituimos el valor de I_E dado en [6.20] se obtiene:

$$I_C = \frac{\alpha_F}{1 - \alpha_F \alpha_R} I_{EO} \left(\exp \frac{q V_{EB}}{kT} - 1 \right) - \frac{I_{CO}}{1 - \alpha_F \alpha_R} \left(\exp \frac{q V_{CB}}{kT} - 1 \right) \quad [6.21]$$

Análogamente, si en [6.20] sustituimos I_C por el valor dado en [6.19], resulta:

$$I_E = \frac{I_{EO}}{1 - \alpha_F \alpha_R} \left(\exp \frac{q V_{EB}}{kT} - 1 \right) - \frac{\alpha_R}{1 - \alpha_F \alpha_R} I_{CO} \left(\exp \frac{q V_{CB}}{kT} - 1 \right) \quad [6.22]$$

Este par de ecuaciones es conocido como *ecuaciones de Ebers-Moll*, y son utilizadas frecuentemente en programas de cálculo de simulación de circuitos, como el programa SPICE³. En este sentido, conviene recordar que los factores α_F y α_R , así como las corrientes I_{EO} e I_{CO} , están determinados por las características del transistor (tipo de semiconductor, dopaje, anchura de la base, etc.) y por tanto su valor puede ser calculado a partir de estos datos. De hecho, es fácil demostrar que no todos estos parámetros son independientes entre sí, ya que están relacionados a través de la ecuación:

$$\alpha_F I_{EO} = \alpha_R I_{CO} \quad [6.23]$$

Las ecuaciones anteriores también se pueden escribir de forma más simplificada:

$$I_C = \alpha_F I_{ES} [\exp (q V_{EB} / kT) - 1] - I_{CS} [\exp (q V_{CB} / kT) - 1] \quad [6.24]$$

$$I_E = I_{ES} [\exp (q V_{EB} / kT) - 1] - \alpha_R I_{CS} [\exp (q V_{CB} / kT) - 1] \quad [6.25]$$

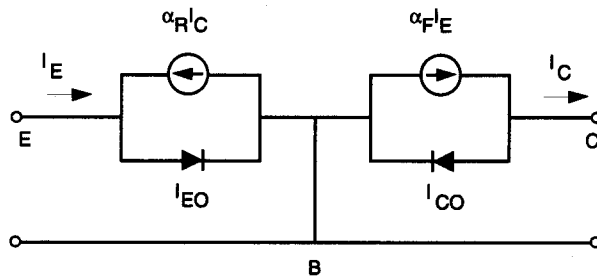
donde los coeficientes I_{ES} e I_{CS} se denominan *corrientes de saturación de emisor y colector*, respectivamente, y están relacionadas con las corrientes I_{EO} e I_{CO} a través de las ecuaciones:

$$I_{ES} = I_{EO} / (1 - \alpha_F \alpha_R) \quad [6.26]$$

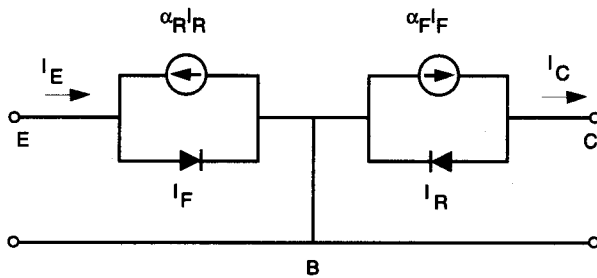
$$I_{CS} = I_{CO} / (1 - \alpha_F \alpha_R) \quad [6.27]$$

Las ecs. [6.24] y [6.25] permiten establecer un modelo alternativo de circuito equivalente del transistor, mostrado en la fig. 6.7b. La similitud con el circuito de la fig. 6.7a es evidente, ex-

³ **Nota:** Acrónimo inglés procedente de "Simulation Program with Integrated Circuit Emphasis". Este programa y otros similares son muy utilizados en los centros de investigación dedicados al diseño de circuitos integrados mediante ordenador.

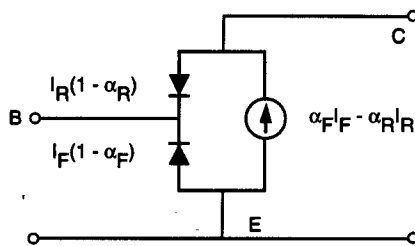


a)



b)

$$\begin{cases} I_F = I_{ES} [\exp(qV_{EB}/kT) - 1] \\ I_R = I_{CS} [\exp(qV_{BC}/kT) - 1] \end{cases}$$



c)

Fig. 6.7. Modelo de Ebers-Moll para el circuito equivalente del transistor, utilizando para las fuentes de corriente un valor proporcional a la corriente en los terminales de emisor y colector (a) o bien proporcional a la corriente en las uniones de emisor y colector (b). El circuito representado en (c) es una variante del caso anterior, utilizado sobre todo para la configuración de emisor común.

cepto que ahora las fuentes de corriente tienen un valor proporcional a la corriente que circula en cada una de las uniones. El circuito de la fig. 6.7a se utiliza muy a menudo para la configuración de base común. Para la configuración de emisor común se pueden usar indistintamente, bien sea el circuito mostrado en la fig. 6.7b, o bien el de la fig 6.7c, cuya equivalencia con el anterior se deja como ejercicio para el lector.

6.3.2. Curvas I-V para la configuración de base común

Las relaciones [6.24] y [6.25] muestran que las corrientes en cada uno de los terminales del transistor dependen directamente de las tensiones aplicadas a cada una de las dos uniones que forman el transistor. Con objeto de tener una idea clara del comportamiento del transistor cuando se utiliza en circuitos amplificadores de una sola etapa es conveniente hacer una representación gráfica de la variación de la las corrientes en el dispositivo frente a las tensiones aplicadas. Para ello recurriremos de nuevo a las dos configuraciones más utilizadas del transistor en circuitos amplificadores, la de base común y la de emisor común, representadas en la figs. 6.6a y 6.6b, respectivamente. Comencemos primero con la configuración de base común haciendo una discusión por separado de la variación de la corriente en los terminales de entrada y salida del transistor.

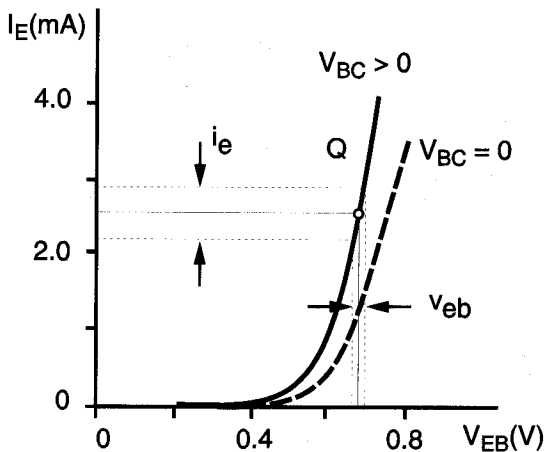


Fig. 6.8. Características de entrada de un transistor pnp en la configuración de base común. Se muestra la variación de la corriente en un punto de funcionamiento Q cuando el transistor opera con señales pequeñas de voltaje.

i) Circuito de entrada (terminales emisor-base):

Las curvas características $I - V$ de un transistor pnp en el circuito de entrada para la configuración de base común vienen representadas en la fig. 6.8. Estas curvas ilustran la variación típica de la corriente de emisor, I_E , en función de la tensión de polarización en la unión de emisor, V_{EB} , usando como parámetro diferentes valores de la tensión entre base y colector, V_{BC} . Es evidente que cuando $V_{EB} > 0$ el comportamiento del transistor visto desde los terminales de emisor y base se acerca mucho al de una unión p^+-n polarizada en directo y de hecho muy a menudo se utiliza el transistor como un simple diodo rectificador, cortocircuitando la unión de colector ($V_{BC} = 0$). De la ec. [6.25] se deduce que siempre que la unión de colector esté polarizada en inverso ($V_{BC} > 0$) la corriente de emisor tiene una variación cuasi-exponencial con V_{EB} y con un valor prácticamente independiente de V_{BC} , ya que el segundo término de la ec. [6.25] es muy pequeño frente al primero. Por esta razón, en circuitos prácticos, la tensión V_{EB} suele ser muy pequeña, alrededor de 0,6 - 0,7 V para transistores de Si, es decir, próxima a la tensión umbral de la unión p-n, ya que a partir de esta tensión la corriente I_E toma valores elevados. Hay que notar en la fig. 6.8 que para $V_{BC} = 0$ la corriente I_E disminuye ligeramente sobre el valor correspondiente al caso anterior, ya que desaparece la contribución del segundo término de la ec. [6.25]. Esta situación también se puede interpretar por la desaparición del efecto de drenaje que ejecuta el colector para los huecos que proceden del emisor cuando la unión de colector se polariza en inversa.

ii) Circuito de salida (terminales base-colector):

En el circuito de salida de la configuración de base común la curva característica $I - V$ del transistor representa la variación de la corriente de colector, I_C , frente a la tensión aplicada en los terminales base-colector, V_{BC} . Como parámetro se podría tomar la tensión entre emisor y base, V_{EB} , del circuito de entrada. Hay que considerar, sin embargo, que en la región activa la intensidad de colector tiene una dependencia lineal con la corriente de emisor, I_E (ec. 6.16), por lo que es preferible utilizar esta corriente como parámetro. En la fig. 6.9 se da una familia de curvas que representa la variación de I_C para diferentes valores de V_{BC} tomando I_E como parámetro para un transistor típico del tipo pnp. Las curvas se suelen trazar de forma que el cambio de I_E de una a otra sea constante (1 mA en el caso de la fig. 6.9). Se aprecia en estas curvas que en la **región activa** ($V_{BC} > 0$) la corriente de colector, I_C , coincide aproximadamente con el valor de la corriente de emisor (recuérdese que $\alpha_{dc} = \alpha_F \approx 1$) y además es prácticamente independiente de la tensión base/colector dentro de un amplio rango de voltajes (ec. 6.24). Existe sin embargo una pequeña corriente, $I_{CBO} = I_{CO}$, incluso cuando la corriente de emisor es cero (emisor en circuito abierto) que corresponde a la corriente de minoritarios, o corriente de fugas, en la unión de colector polarizada en inverso. Para reducir completamente a cero esta corriente de colector sería necesario polarizar también la unión de emisor con una pequeña tensión en inversa. El transistor funciona entonces en la denominada **región de corte** en la que prácticamente no circula ninguna corriente a través de él. En esta situación el circui-

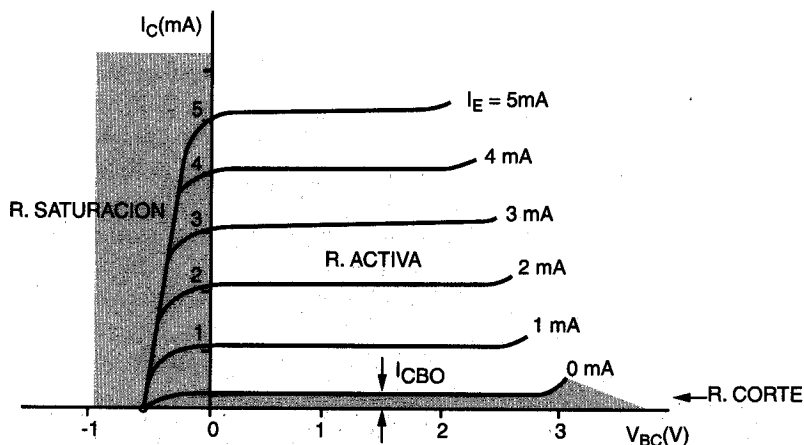


Fig. 6.9. Curvas características I-V de salida de un transistor pnp en la configuración de base común.

to equivalente del transistor se reduce prácticamente a dos diodos conectados en oposición y polarizados en inversa cada uno de ellos.

Por contra, en la **región de saturación**, es decir, para $V_{BC} < 0$, las dos uniones que forman el transistor están polarizadas en directo por lo que desde el emisor y colector se inyectan sendas corrientes en sentido opuesto hacia la base. En estas circunstancias, la corriente que nace en el colector se opone a la corriente principal de huecos procedentes del emisor produciendo una disminución global de la corriente I_C , tal como se observa en la fig. 6.9 (nótese que en la ec. [6.24] la corriente I_C viene ahora dada por la suma de dos términos contrapuestos). Si la polarización V_{BC} se hace negativa en unas décimas de voltio, es decir, con un valor similar a la polarización de la unión de emisor, la corriente I_C se reduce completamente a cero (ec. 6.21). Las curvas de I_C convergen entonces en un punto situado en la región negativa del eje de abscisas. (Véase problema 6.2).

6.3.3. Curvas I-V para la configuración de emisor común

i) Circuito de entrada (terminales base-emisor):

Las curvas I - V del circuito de entrada en la configuración de emisor común corresponden a la variación de I_B con V_{EB} , tomando como parámetro la tensión V_{EC} del circuito de salida. Estas variables (I_B y V_{EB}) no aparecen de forma explícita en las ecuaciones de Ebers-Moll

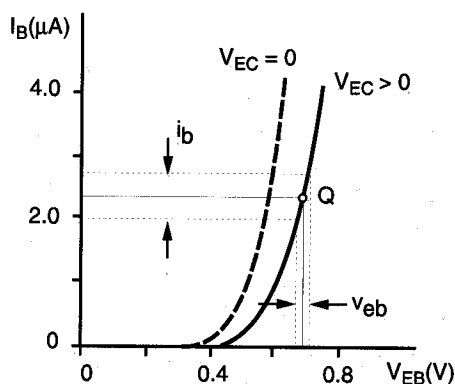


Fig. 6.10. Características I-V de entrada para un transistor pnp en la configuración de emisor común. Se muestra la variación de la corriente en un punto de funcionamiento, Q, cuando el transistor opera con señales pequeñas de voltaje.

por lo que la discusión se hará en términos de los resultados obtenidos para la configuración de base común.

En la fig. 6.10 se ha representado la variación típica de la corriente I_B en función de V_{EB} para los casos de $V_{EC} > 0$, es decir, con una tensión aplicada entre emisor y colector de forma que el transistor se encuentre en la región activa y de $V_{EC} = 0$, colector en corto con el emisor. Cuando el transistor opera en la región activa con $V_{EC} > 0$, la corriente de base es prácticamente independiente de la tensión V_{EC} , por lo que la variación de I_B frente a V_{EB} puede representarse por una curva única (curva continua en la fig. 6.10). Obsérvese que la variación de I_B en este caso es muy similar a la de un diodo polarizado en directo, ya que el comportamiento de unión de emisor es muy similar al de una unión p^+-n . La analogía con las correspondientes curvas de la fig. 6.8 para la corriente de emisor es evidente, aunque al ser $I_B \ll I_E$ el diodo opera en una región mucho más próxima al origen, y por tanto con una resistencia dinámica mayor. En circuitos prácticos, el transistor suele estar polarizado con una tensión V_{EB} pequeña (entre 0,6 y 0,7 V para el Si), es decir, ligeramente superior a la tensión umbral de la unión p-n.

La curva para $V_{EC} = 0$ de la fig. 6.10 (curva a trazos) implica que el emisor y el colector están polarizados al mismo potencial respecto de la base, formando las uniones de emisor y de colector dos diodos en paralelo polarizados en directo. Por tanto la variación de I_B con V_{EB} es en este caso similar a la de un diodo, aunque lógicamente con un valor más elevado de la corriente. Es interesante señalar que la curva correspondiente a $V_{EC} > 0$ no pasa

por el origen, de forma que para $V_{EB} = 0$ la corriente de base toma un valor negativo pequeño. Efectivamente, para $V_{EB} = 0$ no puede haber flujo de corriente a través de los terminales de emisor y base ($I_E = 0$). Sin embargo, puede existir una pequeña corriente negativa en el terminal de base, ya que éste se encuentra polarizado positivamente respecto del colector lo cual implica una pequeña corriente en inverso (negativa) que entra por el terminal de base y se dirige hacia el colector.

ii) Circuito de salida (terminales emisor-colector):

En el circuito de salida las características $I - V$ representan la variación de la corriente de colector, I_C , en función de la tensión aplicada entre los terminales de emisor y colector, V_{EC} (fig. 6.11). En este caso, se toma como parámetro la corriente de base, I_B , perteneciente al circuito de entrada, ya que en la región activa existe una dependencia directa entre la corriente de colector y la corriente de base, según se ha señalado anteriormente. Las curvas se trazan de forma que el cambio de I_B de una curva a otra sea constante ($40 \mu A$ en el caso de la fig. 6.11).

En la región **activa**, la corriente de colector I_C es, aproximadamente, independiente de V_{EC} para un valor un valor fijo de I_B ⁴. Para explicar este hecho hay que considerar que V_{EC} se puede descomponer en dos términos, esto es, $V_{EC} = V_{EB} + V_{BC}$. En esta suma, el primer término, V_{EB} , se mantiene constante, ya que esta magnitud está determinada por el valor impuesto para la corriente I_B . Por tanto las variaciones de la tensión de polarización V_{EC} se re-

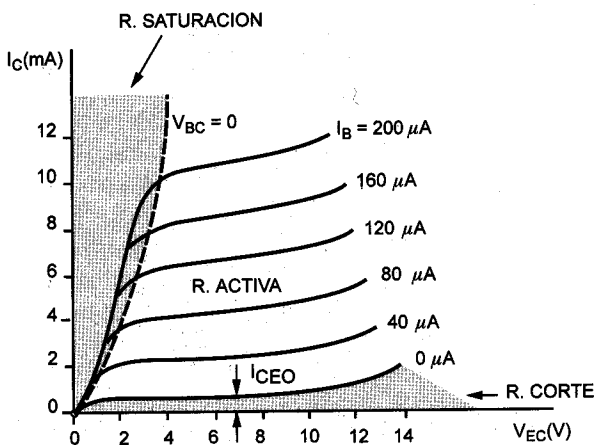


Fig. 6.11. Curvas características $I - V$ de salida de un transistor pnp en la configuración de emisor común.

⁴ **Nota:** En realidad las curvas presentan un ligero incremento de I_C con V_{EC} . Este aumento es debido al efecto de modulación de la base, estudiado en el apartado siguiente.

flejan únicamente en el segundo término, es decir, en la tensión de base/colector, V_{BC} . Según hemos visto para la configuración de base común, la tensión V_{BC} no influye en el valor de la corriente I_C cuando el transistor opera en la región activa.

Hay que notar que, en esta región de operación, la corriente de colector es relativamente alta y en general mucho mayor que I_B , ya que el factor de ganancia β_{dc} suele tener un valor elevado. Además, de acuerdo con la ec. [6.17], para $I_B = 0$ (base en circuito abierto) la corriente de colector, I_C , toma un valor relativamente pequeño, dado por I_{CEO} (véase fig. 6.11). Esta corriente corresponde fundamentalmente a la corriente de huecos en el colector inyectados desde el emisor a través de la base, ya que para un valor dado de V_{EC} la unión de emisor se encuentra siempre con una ligera polarización en directo. Para llevar completamente a cero la corriente de colector sería necesario polarizar la unión de emisor en inverso con una tensión pequeña, tal como se ha señalado anteriormente ($V_{EB} < 0$). El transistor entra entonces en la **región de corte**, caracterizada por una resistencia entre los terminales de emisor y colector elevada, ya que ambas uniones están polarizadas en inverso. En circuitos digitales, cuando el transistor se encuentra operando en esta zona se dice que está en *estado apagado* ("off") o *bajo* ("low").

La **región de saturación** se corresponde con la polarización en directo de la unión de colector ($V_{BC} < 0$), con lo que ambas uniones quedan polarizadas en directo. Según se ha visto más arriba existe entonces una corriente de huecos procedente del colector que se inyecta hacia la base produciendo una disminución de la corriente neta del colector (parte izquierda de las características I-V en la fig. 6.11). En esta situación el transistor opera con una resistencia muy baja entre los terminales de emisor y colector, ya que la corriente I_C puede ser elevada incluso con tensiones V_{EC} bajas. En circuitos digitales la operación en esta zona de corriente elevada se denomina, análogamente, *estado encendido* ("on") o *alto* ("high").

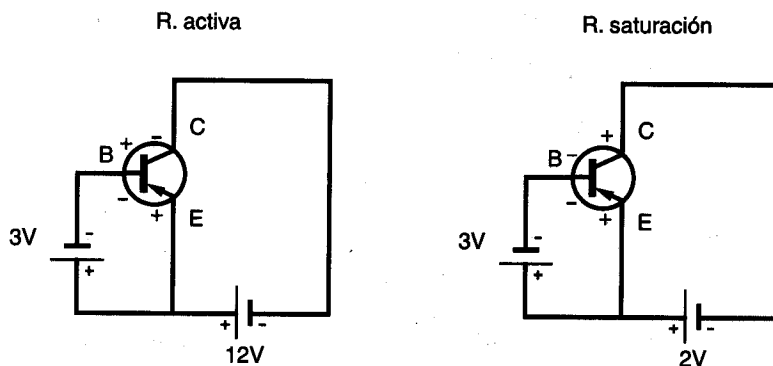


Fig. 6.12. Ejemplos de polarización de un transistor pnp en la región activa (a), y en la región de saturación (b).

Es importante mencionar que, en la región de saturación, la tensión $V_{EC} = V_{EB} + V_{BC}$ es menor que V_{EB} ya que $V_{BC} < 0$. De esta manera, el electrodo de colector se encuentra a un potencial más elevado que el de base, quedando así la unión de colector polarizada en directo. En la fig. 6.12 se ilustra este hecho con un ejemplo comparativo de las tensiones necesarias para polarizar el transistor en la región activa y en la región de saturación. Obsérvese en cada caso las polaridades relativas de los terminales de emisor, colector y base.

Cuando el transistor se utiliza en circuitos amplificadores de tipo analógico para señales alternas normalmente funciona con polarización en un punto de las curvas características situado en la región activa. Por contra, en los circuitos digitales el transistor se polariza de forma que trabaje, bien sea en la región de corte (estado bajo) o bien en la región de saturación (estado alto). El paso de un estado a otro se realiza normalmente cambiando la corriente a través de la base. A este respecto, es curioso observar que la corriente de base puede variar con pequeños cambios en la tensión de polarización de la base, V_{EB} , lo cual a su vez puede dar lugar a cambios importantes en el estado de funcionamiento del transistor. Más adelante, en el capítulo 9, trataremos con más detalle estos aspectos.

6.4. EFECTO DE MODULACION DE LA ANCHURA DE LA BASE

Hasta ahora nos hemos limitado a describir las características del transistor en la región activa a partir de un comportamiento ideal que prevé, entre otras cosas, una corriente de colector prácticamente independiente de la tensión base/colector. Existen, sin embargo, desviaciones importantes de este comportamiento que se reflejan sobre todo en un ligero aumento de la corriente de colector con la tensión base/colector (fig. 6.9) y también en un aumento más acusado con la tensión emisor/colector (fig. 6.11), así como en la existencia de una tensión crítica a partir de la cual la corriente inicia un aumento muy elevado.

Estas desviaciones del comportamiento ideal están ocasionadas por numerosos efectos, quizás el más importante de ellos es el llamado *efecto de modulación de la base*, también denominado *efecto Early* por sus estudios sobre este problema. Este efecto se debe a que la anchura efectiva de la base, w' , es menor que la anchura real, w , cuando el transistor se polariza en la región activa (fig. 6.13a). La reducción de la anchura de la base está originada por la formación de una zona de carga espacial a ambos lados de la unión de emisor y de la unión de colector (zonas sombreadas en la fig. 6.13a). En condiciones normales de operación, estas zonas no contribuyen de manera significativa a los procesos de recombinación que tienen lugar en la base, ya que en ellas la concentración de carga libre es muy reducida. Por este motivo, la existencia en ambos lados de la base de una zona de carga especial reduce el espesor efectivo para la recombinación. Esto, a su vez, da lugar a un aumento del factor de transporte α_T para los huecos que provienen desde el emisor (ec. 6.3). Lógicamente el efecto de reducción será tanto más notable cuando menor sea la anchura de la base.

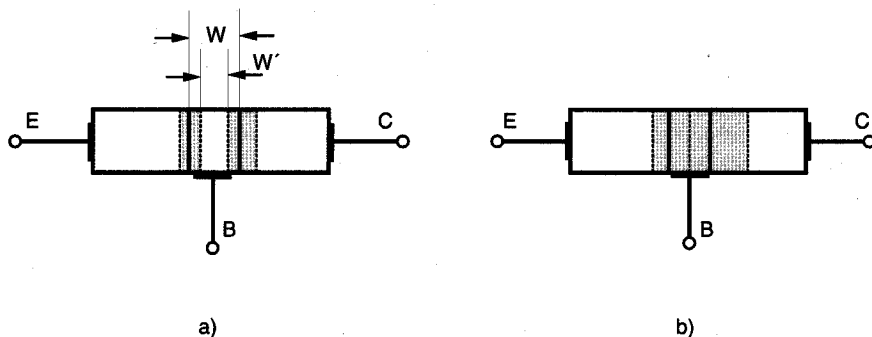


Fig. 6.13. a) Disminución de la anchura de la base en un transistor pnp polarizado en la región activa. b) Efecto de perforación de la base al unirse las regiones de carga espacial de las dos uniones del transistor.

De acuerdo con la expresión [3.11] del capítulo tercero, la anchura de la zona de carga espacial a cada lado de la unión varía inversamente con la concentración de impurezas del semiconductor. Además, ec. [3.16], la anchura total aumenta o disminuye con la tensión aplicada a la unión en inverso o en directo, respectivamente. Así pues, la reducción de la anchura de la base será mayor en la unión de colector, ya que ésta se encuentra polarizada en inversa, generalmente con potenciales más elevados.

El aumento del factor de transporte debido a la disminución de w origina un aumento de la corriente de colector, ya que se inyectan más huecos en el colector. El efecto es particularmente notable en las curvas I-V de salida en la configuración de emisor común, aunque también afecta al resto de las características. En la fig. 6.11 puede observarse cómo, en efecto, existe un ligero aumento de I_C a medida que aumenta V_{EC} en la región activa. En las curvas de salida de la configuración de base común, fig. 6.9, el efecto no es tan aparente, ya que las curvas están trazadas para $I_E = \text{cte}$. La imposición de I_E constante, siendo la corriente de emisor la principal contribución a la corriente de colector, hace que I_C sea prácticamente constante con V_{BC} .

A medida que la tensión inversa aplicada a la unión de colector aumenta, la reducción de la anchura de la base es cada vez mayor de forma que existirá un cierto voltaje para el cual la región de carga espacial de la unión de colector entra en contacto directo con la que existe en la unión de emisor (fig. 6.13b). Este fenómeno, que da lugar a un aumento considerable de la corriente de colector, se denomina *perforación de la base*. Según se aprecia en el esquema

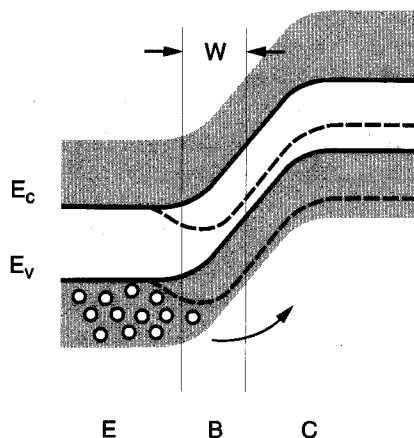


Fig. 6.14. Efecto de la perforación de la base en la curvatura de las bandas de energía de un transistor pnp polarizado en la región activa. Las curvas a trazos indican el estado normal de operación en dicha región.

de bandas de energía de la fig. 6.14 para un transistor pnp, la reducción a cero de la anchura de la base origina una eliminación de la barrera de potencial para la corriente de huecos del emisor. En estas condiciones los huecos pasan directamente de emisor a colector atraídos por la polarización negativa del colector, con el consiguiente aumento de I_C .

Superpuesto al efecto de perforación de base puede existir otro de avalancha de los portadores minoritarios moviéndose a través de la unión de colector polarizada en inverso. Este fenómeno, ya estudiado en el capítulo tercero, aparece a tensiones de polarización inversa elevadas, y da lugar también a un aumento muy apreciable de la corriente a través de la unión. Ambos fenómenos, el de perforación y el de avalancha, pueden coexistir en la región de voltajes elevados de las curvas características, predominando aquel que aparezca a tensiones más bajas.

La fig. 6.15 muestra de forma esquemática el efecto de aumento de la corriente I_C en la región de voltajes elevados de las curvas características de salida como consecuencia de estos dos fenómenos. Generalmente se define un voltaje de ruptura, V_R , como el voltaje umbral a partir del cual se inicia un aumento apreciable de I_C en la curva característica correspondiente a $I_E = 0$ (en la configuración de base común) ó $I_B = 0$ (en la configuración de emisor común). Normalmente V_R es más bajo en este segundo caso, ya que la corriente $I_C = I_{CEO}$ es más

elevada que I_{CBO} , según se señaló anteriormente (apartado 6.2.3). La operación del transistor en la región de ruptura suele ser muy inestable, ya que en esta región se producen cambios considerables de la corriente de colector para variaciones muy pequeñas en el voltaje de polarización de la unión de colector.

Hay que notar además que la región de ruptura frecuentemente cae fuera de la denominada curva de máxima disipación de potencia del circuito de salida del transistor. Esta curva está dada por la ecuación: $P_{\max} = V_{BC} \cdot I_C$, para la configuración de base común, o bien, $P_{\max} = V_{EC} \cdot I_C$, para la configuración de emisor común. Ambas curvas están señaladas por líneas a trazos sobre las curvas características del transistor de la fig. 6.15. El punto operación del transistor en circuitos de amplificación, bien sean analógicos o digitales, ha de elegirse siempre alejado de estas curvas con objeto de evitar un calentamiento excesivo del transistor.

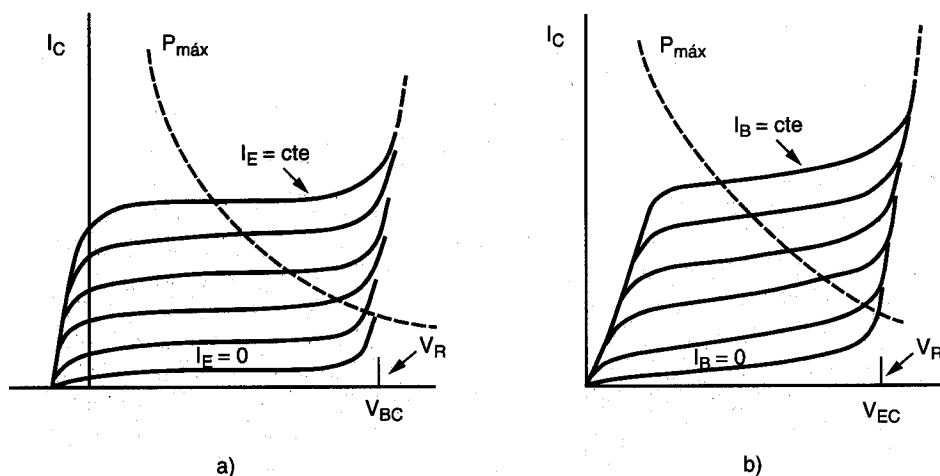


Fig. 6.15. Efecto de la perforación de la base y de avalancha en las curvas I - V de salida para un transistor pnp en la configuración de base común (a) y de emisor común (b). Nótese que debido al efecto de avalancha la corriente I_C alcanza valores muy elevados para valores de V_{BC} ó V_{EC} próximos al potencial de ruptura, V_R .

6.5. COMPORTAMIENTO EN CORRIENTE ALTERNA: CIRCUITO EQUIVALENTE DEL TRANSISTOR PARA SEÑALES PEQUEÑAS

Una de las aplicaciones más frecuentes de los transistores bipolares es la amplificación de señales alternas. En el capítulo 9 se hará una descripción detallada de los circuitos necesarios para efectuar la amplificación. En este apartado se pretende únicamente describir el comportamiento del transistor frente a señales pequeñas de alterna aplicadas sobre los terminales de entrada del transistor.

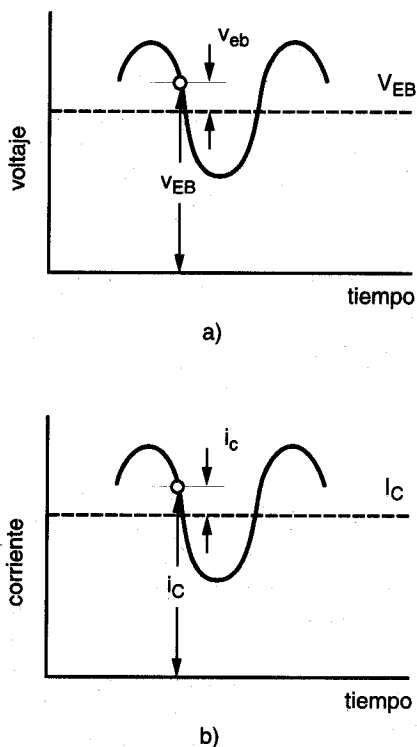


Fig. 6.16. Esquema de la nomenclatura empleada para designar las magnitudes de voltaje (a) y de corriente (b) cuando se opera con señales de alterna.

En los circuitos más sencillos de amplificación el transistor normalmente está polarizado con tensiones en continua, de tal manera que el punto de funcionamiento se sitúa en una zona intermedia de la región activa. Cuando se introduce una señal alterna de tipo sinusoidal en los terminales de entrada del amplificador se produce entonces una variación pequeña sobre los valores de las tensiones de polarización en continua. Así por ejemplo, si el transistor está conectado en la configuración de la base común, la señal de alterna que se pretende amplificar, v_{eb} , se introduce en el transistor a través de los terminales de emisor y base, superpuesta a la tensión de polarización en continua, V_{EB} . La tensión instantánea aplicada a la unión de emisor, v_{EB} , tendrá entonces un valor dado por:

$$v_{EB} = V_{EB} + v_{eb} \quad [6.29a]$$

Esta variación de la tensión entre los terminales de entrada produce una variación pequeña en la corriente de entrada (la corriente de emisor en este caso), también de tipo sinusoidal, cuyo valor señalaremos por i_e . La corriente instantánea, i_E , vendrá dada por:

$$i_E = I_E + i_e \quad [6.29b]$$

Las variaciones de corriente en el circuito de entrada a su vez se traducen en cambios en las correspondientes magnitudes del circuito de salida del transistor produciendo el efecto de amplificación deseado. En lo que sigue, a lo largo de este libro utilizaremos letras minúsculas, v ó i , para distinguir las magnitudes que tienen una variación temporal de tipo sinusoidal de las correspondientes estáticas, las cuales se indicarán con letras mayúsculas. En las señales sinusoidales los subíndices con letras mayúsculas se referirán a los valores totales instantáneos de las magnitudes, mientras que las letras minúsculas en los subíndices se reservarán para indicar la variación de una magnitud en alterna sobre el valor en continua (fig. 6.16).

6.5.1. Circuito equivalente en la configuración de base común

Consideremos un circuito simple como el de la fig. 6.17a, formado por un transistor pnp en la configuración de base común, polarizado en la región activa. Si en el circuito de entrada introducimos una señal alterna, v_{eb} , con un valor de pico mucho menor que la tensión en continua, V_{EB} , el punto de operación del transistor no se modifica sensiblemente dentro de la región activa. Asimismo, supondremos también que la frecuencia de la señal aplicada es lo suficientemente baja como para que las distribuciones de los portadores en cada una de las regiones del transistor puedan seguir instantáneamente las variaciones de la señal aplicada. En estas circunstancias, la tensión v_{eb} origina en el circuito de entrada una señal de corriente, i_e , superpuesta a la corriente continua I_E (no indicada en la figura). La variación de la corriente en el emisor suscita a su vez una variación de la corriente en el colector, i_c , que se superpone al valor continuo I_C . Como veremos más adelante, los circuitos de amplificación se diseñan añadiendo una resistencia R_L en el circuito de salida del transistor, de forma que la señal i_c origine variaciones altas en la caída de potencial a lo largo de la resistencia R_L , produciendo

así una señal de salida amplificada, v_{bc} . Cuando se trata de señales de pequeña magnitud, las relaciones entre las señales de corriente, i_e e i_c y las señales de voltaje, v_{eb} y v_{bc} , presentes en el transistor en los circuitos de entrada y salida, son de tipo lineal, según se demuestra más abajo. Este hecho permite establecer circuitos equivalentes para el transistor relativamente simples, formados por elementos lineales.

Consideremos primero el circuito de entrada del transistor. Según vimos en la sec. 6.3.2, cuando el transistor opera en la región activa existe una relación aproximadamente exponencial entre la corriente I_E y la tensión de polarización en continua, V_{EB} , siendo la corriente I_E prácticamente independiente de la tensión V_{BC} en los terminales de salida. Del mismo modo, cuando existen señales de alterna superpuestas a los valores continuos podre-

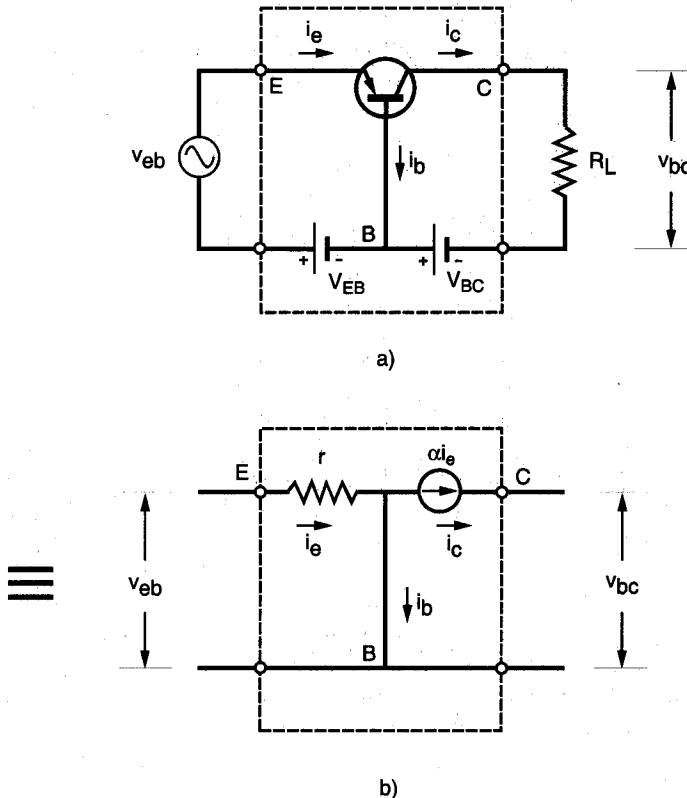


Fig. 6.17. Circuito equivalente (parte inferior) de señales pequeñas en alterna para un transistor pnp operando en la región activa en la configuración de base común (parte superior).

mos establecer para los correspondientes valores instantáneos de la tensión y de la corriente de entrada una relación del mismo tipo que en polarización continua. Sin entrar en detalles de cómo es esta relación para los valores instantáneos, podemos expresarla como:

$$v_{EB} = v_{EB}(i_E) \quad [6.30]$$

Análogamente, para el circuito de salida la corriente continua, I_C , sólo depende de la corriente en el circuito de entrada, I_E , siendo asimismo independiente de la tensión de polarización V_{BC} . Por tanto, si en el circuito existen señales de alterna superpuestas a los valores de continua podemos establecer también para los correspondientes valores instantáneos una relación general del tipo:

$$i_C = i_C(i_E) \quad [6.31]$$

El hecho de trabajar en alterna hace que no sea posible formular una ecuación sencilla para determinar las relaciones de voltajes y las corrientes instantáneas en los circuitos de entrada y salida. Sin embargo, sólo estamos interesados en conocer los valores correspondientes a cambios pequeños producidos por la señal de corriente i_e . Por tanto, utilizando el cálculo diferencial, una pequeña variación en i_E dada por $\Delta i_E = i_e$ produce cambios en v_{EB} y en i_C dados por:

$$v_{eb} = \Delta v_{EB} = \left. \frac{\partial v_{EB}}{\partial i_E} \right|_{V_{BC}} \Delta i_E = r i_e \quad [6.32]$$

$$i_c = \Delta i_C = \left. \frac{\partial i_C}{\partial i_E} \right|_{V_{BC}} \Delta i_E = \alpha i_e \quad [6.33]$$

donde $r = (\partial v_{EB} / \partial i_E)_{V_{BC}}$ corresponde a la resistencia dinámica de la unión de emisor, dada por la inversa de la pendiente en el punto de operación, Q, de las curvas I-V de entrada (fig. 6.8). Asimismo, el coeficiente $\alpha = (\partial i_C / \partial i_E)_{V_{BC}}$ coincide con la definición dada anteriormente para el factor de ganancia en corriente para señales de alterna (ec. 6.8). En circuitos amplificadores, con el transistor operando en la región activa, tanto el coeficiente α como la resistencia r tienen un valor fijo, independiente de la magnitud de las señales alternas producidas en el circuito. Por tanto, la expresión [6.33] muestra una relación de tipo lineal entre la amplitud de la señal de corriente en el circuito de salida, i_c , y la amplitud de la señal de corriente i_e en el circuito de entrada. La corriente i_e , a su vez está relacionada linealmente con la señal de voltaje v_{eb} aplicada entre los terminales de emisor y base del transistor a través de la ec. [6.32].

Las ecuaciones [6.32] y [6.33] permiten establecer un circuito equivalente de pequeña señal para el transistor utilizando elementos lineales, tal como se indica en la fig. 6.17b. Este circuito está formado por una resistencia r entre los terminales de entrada, emisor y base, los cuales reciben la señal de voltaje v_{eb} , y una fuente de corriente de valor αi_e en el terminal de

colector en el circuito de salida. Nótese que, para efectos del comportamiento del transistor en alterna, es posible prescindir en el circuito equivalente de los generadores de continua que polarizan el transistor, ya que una vez que r y α son conocidos para una tensión dada de polarización, las señales de alterna pueden considerarse independientes de los valores en continua dentro de un intervalo amplio de operación.

Es importante destacar que el circuito equivalente de pequeña señal, resultante de este análisis es idéntico al que se obtiene a partir del modelo de Ebers-Moll para la región activa (con $I_R = 0$), sustituyendo el diodo de la unión de emisor por su resistencia r en el punto de operación (fig. 6.7b).

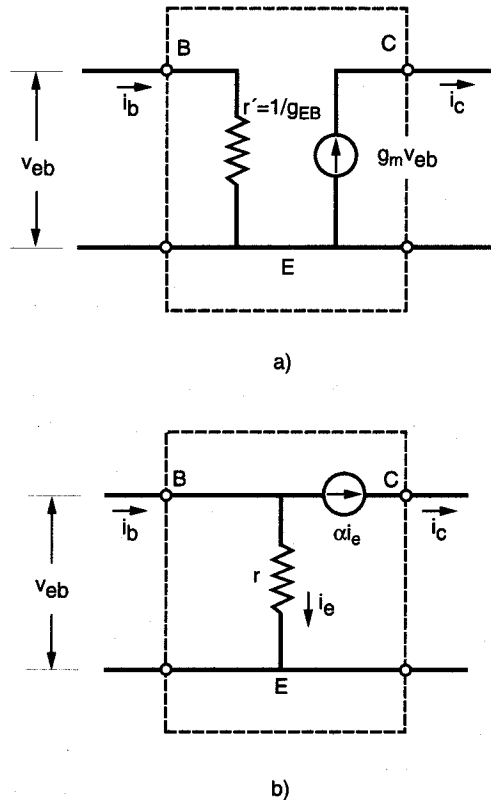


Fig. 6.18. a) Circuito equivalente de pequeña señal para un transistor pnp operando en la región activa en la configuración de emisor común. b) Modelo alternativo de circuito equivalente para la misma configuración, obtenido a partir del circuito utilizado en la configuración de base común.

6.5.2. Circuito equivalente en la configuración de emisor común

Para la configuración de emisor común se puede establecer también un circuito equivalente de pequeña señal similar al de base común. Supondremos ahora que los valores instantáneos de las corrientes de base y de colector, i_B e i_C , son en primera aproximación prácticamente independientes de la tensión v_{EC} . Así pues, siguiendo las mismas pautas que en el circuito de base común, se pueden establecer las siguientes relaciones para los valores instantáneos de la corriente en los circuitos de entrada y salida, respectivamente:

$$i_B = i_B(v_{EB}) \quad [6.34]$$

$$i_C = i_C(v_{EB}) \quad [6.35]$$

Para pequeñas señales, el cálculo diferencial conduce a:

$$i_b = \Delta i_B = \left. \frac{\partial i_B}{\partial v_{EB}} \right|_{V_{EC}} \Delta v_{EB} = g_{EB} v_{eb} \quad [6.36]$$

$$i_c = \Delta i_C = \left. \frac{\partial i_C}{\partial v_{EB}} \right|_{V_{EC}} \Delta v_{EB} = g_m v_{eb} \quad [6.37]$$

siendo g_{EB} y g_m nuevos parámetros característicos del transistor, denominados *conductancia de entrada* y *transconductancia*, respectivamente. De acuerdo con las ecs. [6.36] y [6.37], su valor viene dado por:

$$g_{EB} = \left. \frac{\partial i_B}{\partial v_{EB}} \right|_{V_{EC}} \quad [6.38]$$

$$g_m = \left. \frac{\partial i_C}{\partial v_{EB}} \right|_{V_{EC}} \quad [6.39]$$

Las relaciones anteriores indican que la señal de corriente i_b en el circuito de entrada se puede obtener a partir del voltaje v_{eb} aplicado sobre una resistencia de valor $r' = 1 / g_{EB}$. La resistencia r' corresponde a la resistencia dinámica de la unión de emisor en el punto de operación del transistor, y viene dada por la inversa de la pendiente de las curvas $I - V$ de entrada en el punto de operación Q (fig. 6.10). Asimismo, se puede considerar que en el circuito de salida existe también una señal de corriente, i_c , de valor proporcional a la señal de entrada, esto es: $i_c = g_m v_{eb}$. De acuerdo con este comportamiento, el circuito equivalente de un transistor pnp en la configuración de emisor común viene representado en la fig. 6.18a. Según

se muestra, el circuito contiene la resistencia r' , y el generador de corriente, $g_m v_{eb}$, en los circuitos de entrada y salida, respectivamente. Al igual que en el esquema de la fig. 6.17, se ha prescindido también en este circuito equivalente de los generadores de tensión en continua. Es fácil demostrar, también en este caso, que el circuito de la fig. 6.18 se puede obtener a partir del circuito de la fig. 6.7c resultante del modelo de Ebers-Moll para la región activa (véase problema 6.11).

Dado que la formulación matemática del circuito equivalente es completamente general, es posible extender el circuito equivalente de base común, fig. 6.17, a la configuración de emisor común. Resulta entonces para esta configuración el circuito de la fig. 6.18b, el cual se puede considerar un modelo alternativo al de la fig. 6.18a, pero con elementos diferentes. Aunque este modelo no es muy conveniente, ya que la fuente de corriente está determinada por una variable que no pertenece al circuito de entrada, resulta útil en el análisis de algunos circuitos amplificadores.

En apartados anteriores hemos visto que la corriente en el circuito de entrada de la configuración de emisor común (corriente de base) es mucho menor que la correspondiente al circuito de base común (corriente de emisor). Algo similar ocurre en los circuitos equivalentes de pequeña señal, lo cual se refleja en que la resistencia r' del circuito de entrada de la configuración de emisor común es mucho mayor que la resistencia r para el de base común. Matemáticamente se puede obtener una relación entre ambas resistencias a través de las ecuaciones [6.8] y [6.10], es decir:

$$i_c = \alpha i_e$$

y

$$i_c = \beta i_b$$

Por tanto

$$i_e = (\beta / \alpha) i_b = i_b / (1 - \alpha) \quad [6.40]$$

En la última igualdad se ha hecho uso de la ec. [6.14]. A partir de estas ecuaciones tendremos:

$$r' = \frac{v_{eb}}{i_b} = \frac{r i_e}{i_b} = \frac{1}{1 - \alpha} r = (\beta + 1) r \approx \beta r \quad [6.41]$$

Conforme se había indicado, este resultado muestra que en el circuito equivalente de pequeña señal la resistencia de entrada para la configuración de emisor común es aproximadamente β veces más elevada que la correspondiente resistencia de la configuración de base común.

Los circuitos equivalentes de pequeña señal resultan de mucha utilidad práctica cuando se hace el análisis del comportamiento en alterna de circuitos amplificadores complejos.

De hecho, los parámetros α , r , g_m , etc., característicos del transistor, están relacionados directamente con un conjunto más amplio conocido como *parámetros híbridos* del transistor. Estos parámetros se incluyen normalmente en las especificaciones dadas por el fabricante para un punto determinado de funcionamiento en la región activa, con objeto de tener una información rápida del comportamiento del transistor.

6.6. COMPORTAMIENTO DEL TRANSISTOR FRENTE A PULSOS DE CORRIENTE (*)

Muy a menudo el transistor bipolar es utilizado en circuitos digitales como interruptor o puerta lógica. En este tipo de aplicaciones, cuando el transistor está conectado en la configuración de emisor común, la resistencia entre los terminales de colector y emisor pasa de un valor alto, o *estado apagado* (con el punto de operación en la región de corte, es decir, de corriente baja) a un valor bajo, o *estado encendido* (punto de operación en la región de saturación, es decir, de corriente elevada). Esto se consigue mediante la aplicación de un voltaje adecuado en la unión de emisor. En este tipo de aplicaciones interesa conocer el comportamiento del transistor ante pulsos rápidos de tensión (o corriente) aplicados en los terminales que polarizan la unión de emisor con objeto de determinar el tiempo de interrupción, esto es, el tiempo que tarda el transistor para pasar del estado apagado al estado encendido o viceversa. Como veremos a continuación el tiempo de interrupción está controlado fundamentalmente por la variación de carga acumulada en la base por los portadores minoritarios (huecos, para un transistor pnp).

Consideremos el caso de un transistor pnp que se encuentra inicialmente en estado apagado. En esta situación las dos uniones se encuentran polarizadas en inversa de forma que las corrientes de base y emisor son prácticamente nulas (fig. 6.19a). La carga debida a los huecos minoritarios en la base es también nula y el punto de operación (P_1) se halla en la región de corte (fig. 6.19b). Supongamos que en el instante $t=0$ polarizamos la unión de emisor de forma que la corriente de base pasa de cero a un valor I_B elevado, suficiente para trasladar el punto de operación a la región de saturación (P_2). La carga en la base debida a los huecos procedentes del emisor (portadores minoritarios) no pasa inmediatamente al valor correspondiente a la región de saturación. Ello es debido a la capacidad de difusión de la unión de emisor, la cual da lugar a que la carga acumulada en la base se acerque a su valor estacionario, en el punto P_2 , mediante una variación de carácter exponencial. Se puede demostrar que en este proceso de carga la constante de tiempo coincide con el tiempo de vida media de los huecos, τ_{th} , definido en el apartado 2.6.1. En estas condiciones, la corriente de colector tampoco puede aumentar instantáneamente al valor de I_C correspondiente a la saturación. Un cálculo detallado de la corriente permite demostrar que durante la subida la corriente de colector viene dada en cada instante por el cociente $Q(t)/t_B$, siendo $Q(t)$ el valor de la carga de huecos minoritarios en la base y t_B el tiempo de tránsito de los huecos a través de la base. Así pues,

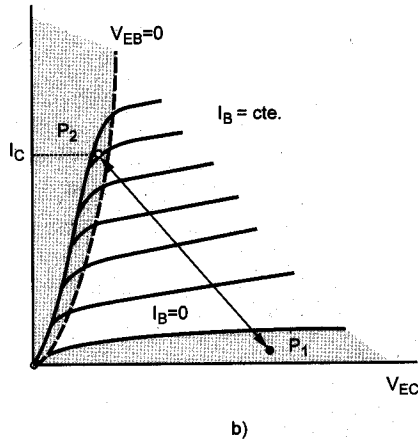
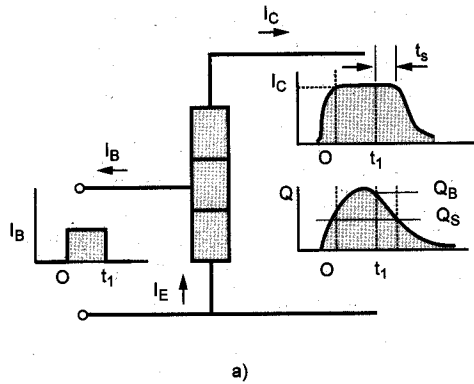


Fig. 6.19. a) Esquema de la variación de los pulsos de corriente en los circuitos de entrada y salida de un transistor pnp funcionando como interruptor en la configuración de base común. b) Variación del punto de operación en las curvas características del circuito de salida del transistor.

en el transitorio de subida la corriente de colector aumenta proporcionalmente a la carga de la base, siguiendo también una ley de carácter exponencial hasta alcanzar un valor I_C . Por tanto, el tiempo de subida se corresponde con el tiempo necesario para trasladar el punto de funcionamiento de P_1 a P_2 pasando por la región activa, y este tiempo está determinado esencialmente por el tiempo de vida media de los portadores minoritarios en la base.

Una vez que el transistor alcanza el punto P_2 en la región de saturación, la corriente I_C permanece prácticamente constante, siempre que I_B sea también constante. No ocurre así con

la carga acumulada en la base cuyo valor máximo, Q_B , viene dado por la ec. [3.50] es decir: $Q_B = I_B \tau_h$. Este valor siempre será mayor que la carga Q_s acumulada durante el tiempo de subida.

Precisamente, esta carga acumulada en exceso en la región de base es la que hace que el transistor se mantenga en la región de saturación durante un cierto tiempo, t_s , una vez que la corriente I_B se reduce instantáneamente a cero cuando el transistor pasa de nuevo al estado de apagado. El tiempo t_s , denominado *tiempo de almacenamiento* es el necesario para que la carga Q_B disminuya hasta Q_s . En este período la corriente de colector se mantiene relativamente constante, ya que la distribución de huecos en la región de base, que es el factor determinante de I_C , se mantiene prácticamente inalterada. Una vez que $Q(t) = Q_s$, la corriente de colector inicia una caída rápida con una cinética similar a la del proceso de subida, es decir, siguiendo una ley de tipo exponencial con una constante de tiempo dada por τ_h .

Todos estos resultados indican que la velocidad de respuesta de un transistor pnp a impulsos rápidos de corriente está controlada por el tiempo de vida media de los huecos minoritarios en la base en los procesos de recombinación. En circuitos digitales en los que se requiere una velocidad alta de conmutación interesa reducir τ_h todo lo que sea posible. Para lograr este objetivo se recurre, durante la fabricación del transistor, a métodos que permitan introducir centros de recombinación en la región de base, según se comentó en el apartado 2.6. Estos centros de recombinación, que están constituidos por estados energéticos localizados en las proximidades del centro de la banda prohibida del semiconductor, aumentan la velocidad de recombinación de los portadores.

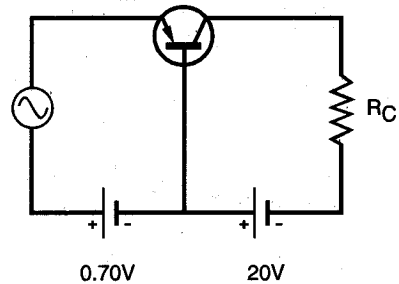
CUESTIONES Y PROBLEMAS

- 6.1 Dibujar esquemáticamente las características I-V de salida, en las cuatro regiones de funcionamiento, de un transistor conectado en la configuración de emisor común. Indicar en cada caso el diagrama de bandas de energía del transistor, discutiendo el movimiento de los portadores.
- 6.2 Explicar porqué la familia de curvas de salida en la configuración de base común (fig. 6.9) para efectos prácticos converge en un punto único situado en la parte negativa del eje de abscisas.
- 6.3 Diseñar un circuito simple para polarizar un transistor de Si en la región activa, en las configuraciones de emisor y de base común, imponiendo la condición de que la corriente de emisor sea de 2.0 mA (utilícese para α_{dc} el valor de 0.98).

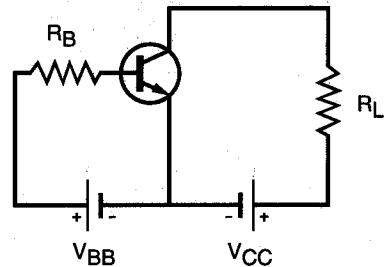
6.4 Un transistor de silicio de tipo pnp tiene el emisor dopado con $N_a = 10^{20} \text{ cm}^{-3}$ y la base con $N_d = 2 \times 10^{17} \text{ cm}^{-3}$. Si la anchura de la base es de $4 \mu\text{m}$, determinar la eficiencia del emisor y el coeficiente de transporte de la base (utilícese para L_h el resultado del problema 3.2).

6.5 En un transistor se tiene: $I_C = 7.606 \text{ mA}$, $I_B = 70 \mu\text{A}$ y $I_{CBO} = 6 \mu\text{A}$. Calcular los valores de α_{dc} , β_{dc} , e I_E .

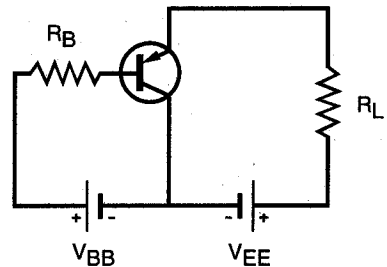
6.6 Suponiendo que el transistor del circuito adjunto tiene la característica I-V de entrada similar a la curva de la fig. 6.8, y que el generador del circuito de entrada proporciona un voltaje de pico de 0.025 V , calcular de forma aproximada el factor de amplificación del voltaje a través de la resistencia $R_C = 5000 \text{ ohm}$ en el circuito de colector (utilícese para α el valor de 0.98).



6.7 El transistor de silicio de la figura tiene $\beta_{dc} = 50$. Calcular los voltajes y corrientes en el transistor, siendo $V_{BB} = 3.1 \text{ V}$; $V_{CC} = 15 \text{ V}$; $R_B = 20 \text{ K ohm}$; $R_L = 1.5 \text{ K ohm}$. (Nota: para los transistores de silicio V_{BE} vale 0.6 V , aproximadamente).



6.8 En la configuración de colector común de la figura adjunta, el transistor es de silicio con $\beta_{dc} = 50$. Calcular los voltajes y corrientes en el transistor, siendo $V_{BB} = 3 \text{ V}$, $V_{EE} = 8 \text{ V}$, $R_B = 8 \text{ K ohm}$ y $R_L = 200 \text{ ohm}$. (tómese $V_{EB} = 0.6 \text{ V}$).



6.9 Para el transistor pnp en la configuración de emisor común, cuyas características de salida se representan en fig. 6.11: a) Calcular de forma aproximada su resistencia de salida para el punto de funcionamiento de la región activa especificado por $V_{EC} = 6 \text{ V}$, $I_B = 120 \mu\text{A}$. ¿Cómo se compara este valor con el que ofrece el transistor en base común?, b) Hallar aproximadamente la ganancia de corriente del transistor en el mismo punto.

- 6.10** Explicar por qué en los modelos de circuito equivalente de pequeña señal, figs. 6.17 y 6.18, se supone que el transistor se comporta como una fuente de corriente constante en lo que se refiere al circuito de salida.
- 6.11** Mostrar la equivalencia del circuito de pequeña señal del transistor en la configuración de emisor común (fig. 6.18a) con el obtenido a partir del modelo de Ebers-Moll para la región activa (fig. 6.7c).
- 6.12** Utilizando el modelo de Ebers-Moll, calcular la ecuación de la curva característica I-V en el circuito de entrada y de salida de un transistor pnp en la configuración de emisor común.

CAPITULO VIII

TRANSISTORES DE EFECTO CAMPO (JFET, MESFET Y MOSFET)

Los transistores de efecto de campo o FET¹ se denominan así porque durante su funcionamiento la señal de entrada crea un campo eléctrico que controla el paso de la corriente a través del dispositivo. Estos transistores también se denominan unipolares para distinguirlos de los transistores bipolares de unión (cap. 6) y para destacar el hecho de que sólo un tipo de portadores -electrones o huecos- interviene en su funcionamiento.

Los transistores de efecto campo de unión (JFET) fueron primero propuestos por Schockley en 1952 y su funcionamiento se basa en el control del paso de la corriente por el campo aplicado a la puerta, constituida por una o varias uniones p-n polarizadas en inverso. Los transistores de efecto de campo de unión metal-semiconductor (MESFET), propuestos en 1966, tienen un funcionamiento muy similar al JFET, pero en ellos el control del flujo de corriente se realiza por una unión metal-semiconductor de tipo Schottky (cap. 4). Finalmente, existe también otro tipo de transistores denominados genéricamente MOSFET (metal-óxido-semiconductor), de desarrollo más reciente, en los que el control de la corriente a través del semiconductor se realiza mediante un contacto separado del semiconductor por una capa aislante (normalmente, óxido de silicio). Este tipo de transistores se utiliza preferentemente en la electrónica digital.

En comparación con los transistores bipolares, los FET presentan una impedancia de entrada muy elevada y además consumen muy poca potencia, por lo que su uso se ha extendido sobre todo en

¹ **Nota:** El acrónimo FET proviene del nombre en inglés: "Field Effect Transistor". Asimismo JFET procede del nombre: "Junction Field Effect Transistor".

los circuitos integrados. También encuentran aplicaciones en circuitos de alta frecuencia (microondas), especialmente los MESFET de arseniuro de galio, los cuales tienen un tiempo de respuesta muy rápido debido a la alta movilidad de los electrones en este material.

8.1. EL TRANSISTOR DE EFECTO CAMPO DE UNION (JFET)

8.1.1. Descripción del transistor

Según se muestra en la fig. 8.1a, la estructura básica de un JFET está formada por un semiconductor, de tipo n por ejemplo, con dos contactos óhmicos en sus extremos, uno de ellos S denominado *fuelle o surtidor* y el otro D conocido por el nombre de *drenador o sumidero*. El tercer electrodo G, denominado *puerta*, está constituido por dos regiones de tipo p difundidas a ambos lados de la estructura del semiconductor. Se forma así en el contacto de puerta dos uniones p-n, las cuales están conectadas entre sí y polarizadas en inverso, de forma que la corriente que pasa a través de ellas es prácticamente nula. Generalmente la unión de puerta es del tipo p⁺-n, lo cual significa que la región p de la puerta está mucho más dopada que la región n del semiconductor.

Los JFET utilizados en circuitos integrados normalmente se fabrican siguiendo la *tecnología planar* (véase cap. XIII), según la cual el semiconductor está formado por una capa de carácter n (capa epitaxial) depositada sobre un sustrato de silicio u otro semiconductor, de tipo p. Un área pequeña de la superficie de esta capa epitaxial está difundida con impurezas también de tipo p, y forma, junto con el sustrato de silicio, el contacto de puerta. Los electrodos metálicos de fuente y de sumidero se depositan directamente a ambos lados del contacto superior de puerta. En la fig. 8.1b se presenta un esquema de la geometría de las diferentes zonas y contactos de un JFET utilizado en circuitos integrados, mostrando la *zona activa*, es decir la región donde tiene lugar la acción del transistor. El símbolo de circuitos para el JFET de tipo n se ha representado en la fig. 8.1c, con la flecha de la puerta apuntando hacia dentro para indicar el carácter n del semiconductor. En el caso del JFET de tipo p, la flecha de la puerta tiene la dirección opuesta a la de la figura.

Si el semiconductor es de tipo n, la polarización se hace de forma que la corriente de electrones (portadores mayoritarios) fluya desde el contacto de surtidor, S, al de drenador, D. Esto quiere decir que la tensión $V_{DS} = V_D - V_S$ debe ser positiva. Se dice entonces que el semiconductor funciona como un *canal* de baja resistencia para los electrones, estando limitado el canal por las paredes que forman las dos regiones de carga espacial adyacentes a las uniones p-n de la puerta. Según hemos visto en el capítulo tercero, en las regiones de carga espacial la concentración de carga libre es muy baja y por tanto su resistividad es muy elevada.

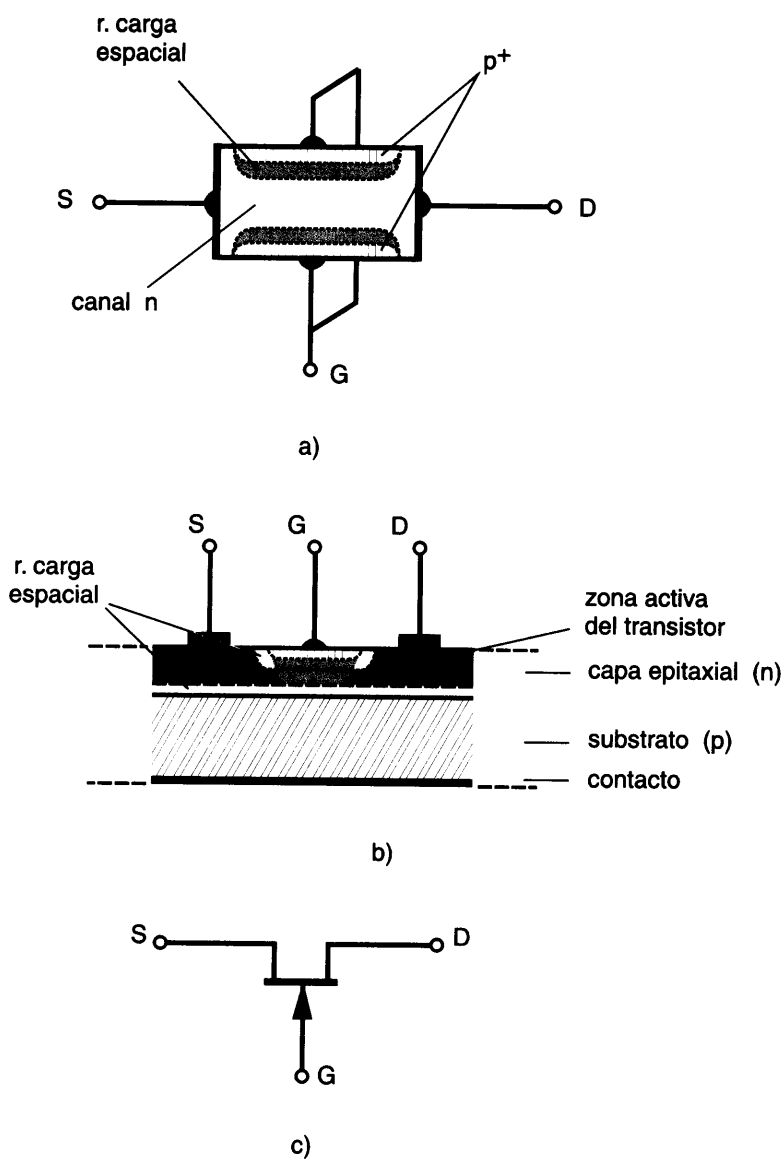


Fig. 8.1. a) Esquema de la estructura de un JFET de canal tipo n. b) Esquema de un transistor JFET de canal n fabricado según la tecnología planar. c) Símbolo de circuitos de un JFET de canal n.

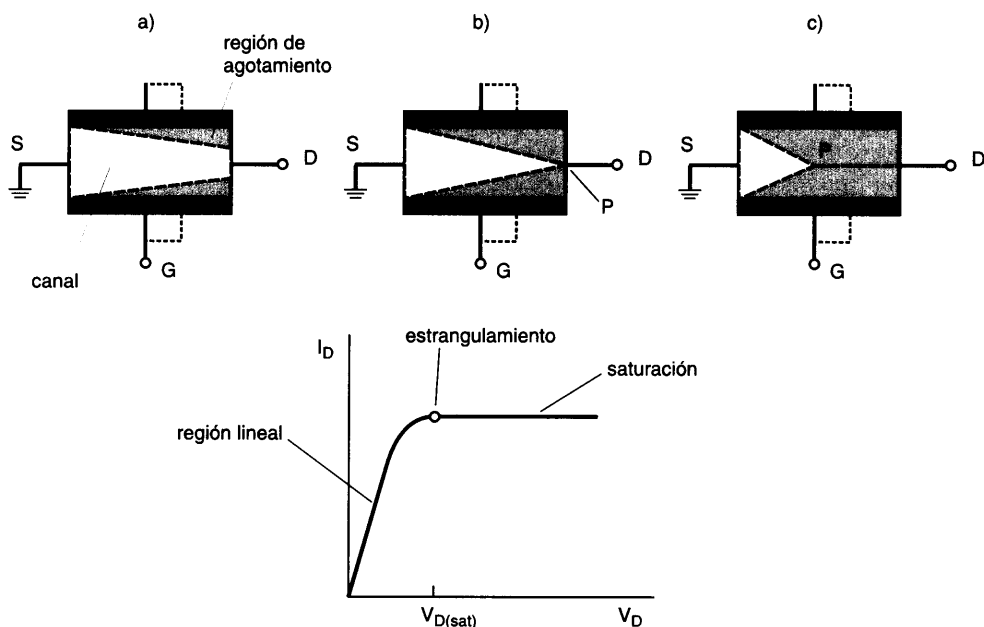


Fig. 8.2 Parte superior: Esquema de la variación de anchura del canal para valores crecientes del voltaje de drenador, V_D , de un transistor JFET: a) Para $V_D \ll V_{D,sat}$. b) $V_D = V_{D,sat}$. c) $V_D \gg V_{D,sat}$. Parte inferior: Representación de la curva de variación de la corriente de drenador, I_D .

8.1.2 Comportamiento cualitativo del JFET.

Supongamos inicialmente que el electrodo de puerta está polarizado al potencial de tierra, esto es $V_G = 0$, y que aumentamos lentamente desde cero la tensión aplicada al drenador, V_D , manteniendo el surtidor a una tensión fija de cero voltios, esto es $V_S = 0$ (potencial de tierra). Para pequeños valores de V_D , la corriente que circula entre el surtidor y el drenador, I_D , debe ser pequeña. Esta corriente es debida al movimiento de electrones desde la fuente al drenador a través del canal. En esta situación se considera que el canal está completamente abierto, comportándose del mismo modo que una resistencia (fig. 8.2a). Así pues, la variación de I_D en función de V_D en el rango de tensiones bajas será prácticamente de tipo lineal. Obsérvese en la fig. 8.2a que el canal semiconductor de tipo n está sometido a un potencial positivo respecto los contactos de puerta (polarizados a una tensión $V_G = 0$) por lo que la unión $p^+ - n$ de los contactos de puerta queda polarizada en inverso. Como consecuencia de esta polarización en inverso, la corriente que circula a través de los contactos de puerta debe ser extremadamente baja. Además, la región de carga espacial, también llamada de agotamiento, que se extien-

de a ambos lados de cada una de las dos uniones tiene una anchura mayor en la región del canal, ya que ésta es la que está menos dopada (región sombreada en la fig. 8.1a). Es importante señalar que **la anchura de la zona de agotamiento es más acusada conforme se avanza hacia el drenador, ya que el potencial $V(x)$ a lo largo del canal semiconductor (región n) es cada vez más positivo respecto de la puerta (región p).**

A medida que se aumenta V_D manteniendo $V_S = 0$, la anchura de la región de carga espacial a ambos lados de la unión $p^+ - n$ de la puerta es cada vez mayor. Este efecto da lugar a una reducción de la anchura del canal, siendo lógicamente el efecto más acusado en la zona del drenador. La reducción gradual de la anchura del canal puede ser tan elevada que puede incluso cerrar el canal en el extremo del drenador. Se origina entonces en esta región un aumento notable de la resistencia del canal, por lo que la pendiente de la curva I_D en función de V_D comienza a decrecer. Cuando el voltaje alcanza un cierto valor crítico, dado por $V_D = V_{D,sat}$ (este valor normalmente es de unos pocos voltios, aunque lógicamente depende de las dimensiones del dispositivo), se llega a la situación indicada en la fig. 8.2b, en la cual se produce el **estrangulamiento** del canal ("pinch-off") y la curva característica $I_D - V_D$ se hace entonces prácticamente horizontal (es decir, resistencia dinámica infinita). A partir de este momento, al aumentar V_D ya no aumenta más la corriente que circula a través del canal. En este rango de tensiones elevadas, conocido como **región de saturación**, el aumento de V_D da lugar a un crecimiento de la longitud de la región del canal que ha sido estrangulada, debido al desplazamiento del punto P (punto donde se produce inicialmente el estrangulamiento) hacia la región de la fuente, fig. 8.2c. Obsérvese que, contrariamente a lo que podría parecer en una primera impresión, al quedar el canal bloqueado la corriente I_D no se reduce a cero, ya que entonces no habría caída de potencial a lo largo del canal y no se llegaría a la condición de estrangulamiento.

Es preciso tener en cuenta además que a medida que el potencial en el drenador se acerca o incluso se hace más elevado que el potencial de estrangulamiento, la caída de potencial a lo largo del semiconductor, $V(x)$, ya no es homogénea. De hecho, cuando se aplica en el drenador un potencial mayor que $V_{D,sat}$ el punto P de estrangulamiento separa dos zonas de diferente resistencia en el semiconductor: entre la fuente y el punto P tenemos el canal semiconductor con resistencia baja, mientras que desde el punto P al drenador la resistencia es muy elevada, ya que corresponde a la región de agotamiento de la unión p-n polarizada en inversa. Es más, el potencial en el punto P de estrangulamiento ha de mantenerse constante e igual al valor de saturación, $V_{D,sat}$. Esto quiere decir que la tensión aplicada, V_D , se reparte de forma no homogénea, produciendo una caída igual a $V_{D,sat}$ en la zona del canal comprendida entre la fuente y el punto P, y una caída igual a $V_D - V_{D,sat}$ en la parte posterior del canal, entre P y el drenador. Para tensiones de drenador mucho más elevadas que $V_{D,sat}$, esta última región es la que absorbe la mayor parte de la tensión aplicada. Este reparto inhomogéneo de la tensión hace que en la región de saturación la corriente de fuente a sumidero sea prácticamente constante ya que está limitada por la caída de tensión entre la fuente y el punto P de estrangulamiento. Un esquema cualitativo de las dos regiones de diferente resistencia que forman el canal en la zona de saturación viene dado en la fig. 8.3.

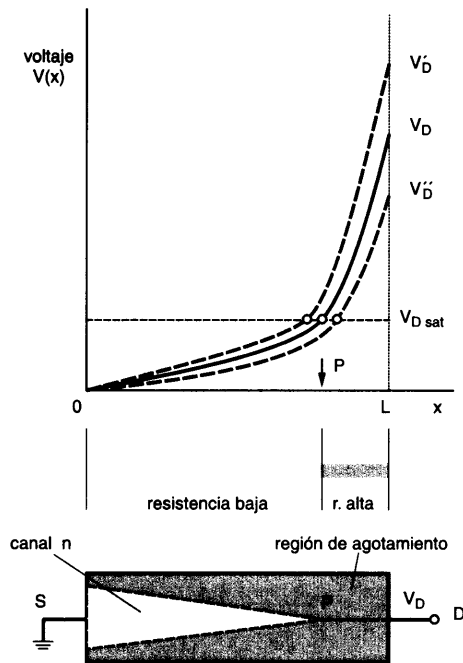


Fig. 8.3. *Parte superior: Variación del voltaje, $V(x)$, a lo largo del canal de un JFET (tipo n) cuando se aplica un voltaje V_D superior al de estrangulamiento (punto de operación en la región de saturación). Las líneas a trazos indican la situación correspondiente a diferentes valores de V_D ($V'_D > V_D > V''_D$). Parte inferior: Esquema de la situación de estrangulamiento del canal en un punto P , en el interior del JFET.*

Aparte de este efecto, existe otra limitación de la corriente que se presenta sobre todo en los transistores con una longitud de canal, L , pequeña. La corriente en este caso puede quedar limitada por la velocidad máxima que alcanzan los portadores en la parte posterior del canal (detrás del punto P), donde el campo eléctrico toma valores muy elevados. Efectivamente, se sabe que para campos relativamente altos, del orden de 10^4 Vcm^{-1} para el silicio, la velocidad de los electrones alcanza un valor de saturación (véase fig. 8.10), ya que la movilidad de los electrones en este rango decrece al aumentar el campo eléctrico aplicado. Este efecto de saturación de la velocidad de los electrones en la zona posterior del canal se puede superponer al ya mencionado de la constancia de la caída de tensión en la parte anterior del canal, produciendo uno u otro una limitación de la corriente en la región de voltajes superiores al voltaje de estrangulamiento.

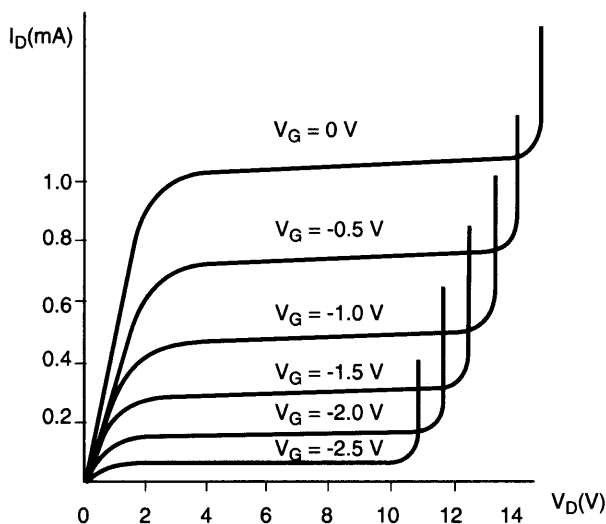


Fig. 8.4. Curvas características I_D - V_D de un JFET típico de canal n, mostrando el aumento abrupto de la corriente en la región de avalancha.

Si en el JFET de canal n polarizado a una tensión de drenador positiva determinada aplicamos ahora tensiones, V_G , negativas en el contacto de puerta, la unión p-n de la puerta queda polarizada en inverso a un potencial más elevado que en el caso anterior (con $V_G = 0$). En esta situación la anchura de la región de carga espacial en cada unión se hace todavía mayor, y a su vez la región del canal se hace más estrecha, por lo que la corriente a través del canal disminuye. En la fig. 8.4 se muestra la familia de curvas características I_D - V_D de un JFET típico de canal n, cada una de ellas correspondiente a un valor de V_G fijo. Nótese que al aumentar el voltaje en inverso, V_G , aplicado a las uniones p-n de la puerta, I_D disminuye como consecuencia del estrechamiento del canal.

En la figura 8.4 se observa también un aumento abrupto de la corriente cuando se alcanza una tensión crítica por encima del valor de saturación. Este aumento se debe a un fenómeno de avalancha de electrones originado en la unión p-n que existe entre la puerta y el canal. Este fenómeno se produce mayoritariamente en la parte posterior del canal, ya que en esta región es donde el diodo tiene una polarización inversa más elevada. Las características de este proceso de ruptura por avalancha en una unión p-n ya fueron señaladas en el capítulo tercero (sec. 3.3.4). Evidentemente, este fenómeno ocurrirá para un valor de V_D tanto menor cuanto más negativo sea V_G .

8.2. CALCULO DE LAS CURVAS CARACTERISTICAS INTENSIDAD-VOLTAJE DEL JFET

Para voltajes de drenador inferiores al de saturación, $V_D < V_{D,sat}$, se puede obtener una relación cuantitativa entre la corriente, I_D , y el voltaje aplicado, V_D , utilizando para el semiconductor el modelo de la fig. 8.1a en dos dimensiones. Para efectuar el cálculo supondremos además (véase fig. 8.5) que: i) se trata de un dispositivo de longitud de canal L y semianchura a , simétrico con respecto a un plano horizontal, ii) las caídas de voltaje en el contacto de fuente (punto de abscisa $x=0$) y del drenador (punto $x=L$) son despreciables, y iii) la variación de las variables electrostáticas en la dirección horizontal (x) es mucho más lenta que en la vertical, lo cual ocurrirá si $L \gg a$ (aproximación del canal gradual).

Consideremos el JFET de canal n tal como se representa en fig. 8.5, es decir antes del estrangulamiento total del canal. Debido al estrechamiento del canal su resistencia es más elevada a medida que se avanza hacia el drenador. Sin embargo, la corriente I_D que fluye a

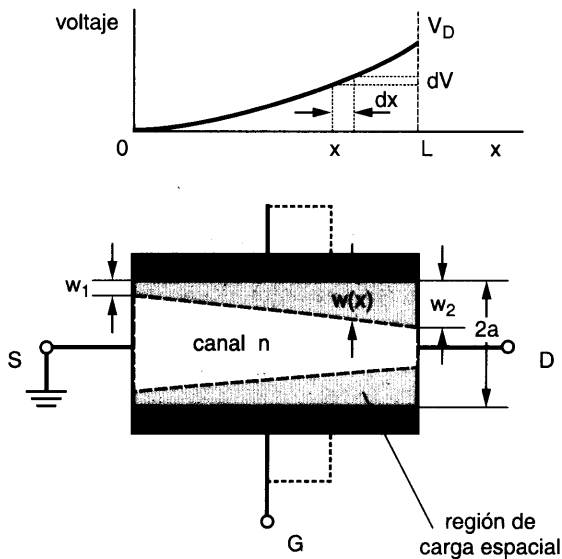


Fig. 8.5. Esquema de la sección transversal de un JFET operando en la región lineal. En la parte superior se muestra la variación del voltaje $V(x)$ a lo largo del canal para un JFET de canal n .

través de cualquier sección transversal, perpendicular a la dirección x , debe ser constante a lo largo del canal. Así pues, para cualquier punto entre $x=0$ y $x=L$ se tendrá:

$$dV = I_D dR \quad [8.1]$$

siendo dR y dV la resistencia y la caída de potencial correspondientes a una longitud dx del canal. Lógicamente, el valor de dR depende de la anchura del canal en el punto considerado y puede calcularse a través de la relación:

$$dR = \rho (dx / S)$$

donde $S = 2z [a - w(x)]$ es el área de la sección transversal del canal en el punto x , y ρ la resistividad del semiconductor. La dimensión z corresponde a la anchura del canal en la dirección perpendicular al plano de la figura. Si el semiconductor es de tipo n , con una concentración N_d de impurezas, podemos calcular su resistividad mediante la relación: $1/\rho = q \mu_e N_d$, siendo μ_e la movilidad de los portadores (electrones). Introduciendo los valores de ρ y S en la ecuación anterior tendremos:

$$dR = \frac{dx}{2q \mu_e N_d z [a - w(x)]} \quad [8.2]$$

con lo que sustituyendo en [8.1], resulta,

$$I_D dx = 2z q \mu_e N_d [a - w(x)] dV \quad [8.3]$$

Ahora bien, según el resultado obtenido en ec. [3.47], en una unión abrupta tipo p^+-n la anchura, $w(x)$, de la zona de carga espacial en la región n viene dada por una expresión del tipo:

$$w(x) = \left[\frac{2\epsilon}{q N_d} [V_o + V(x) - V_G] \right]^{1/2} \quad [8.4]$$

ya que el potencial inverso aplicado a través de la unión $p-n$ en el punto de coordenada x es $V_G - V(x)$. En la ecuación anterior, V_o es el denominado potencial barrera o de contacto de la unión de puerta y $V(x)$ es la parte del potencial que cae entre la fuente ($x=0$) y el punto considerado (de coordenada x) como consecuencia del potencial V_D aplicado en el drenador. A partir de la ec.[8.4] se puede establecer para dV :

$$dV = \frac{q N_d}{\epsilon} w dw \quad [8.5]$$

que sustituida en [8.3] nos da:

$$I_D dx = 2z q \mu_e N_d [a - w(x)] \frac{q N_d}{\epsilon} w dw$$

Integrando sobre la longitud L del canal y despejando I_D queda:

$$\begin{aligned} I_D &= \frac{2z \mu_e q^2 N_d^2}{\epsilon L} \int_{w_1}^{w_2} (a - w) w dw = \\ &= \frac{z \mu_e q^2 N_d^2}{\epsilon L} \left[a (w_2^2 - w_1^2) - \frac{2}{3} (w_2^3 - w_1^3) \right] \end{aligned} \quad [8.6]$$

Particularizando para los valores de w_1 , cuando $x=0$, $V=0$, y w_2 , cuando $x=L$, $V=V_D$ y haciendo uso de [8.4] se tiene finalmente:

$$I_D = I_P \left[\frac{V_D}{V_P} - \frac{2}{3} \left(\frac{V_D + V_o - V_G}{V_P} \right)^{3/2} + \frac{2}{3} \left(\frac{V_o - V_G}{V_P} \right)^{3/2} \right] \quad [8.7]$$

donde:

$$I_P = \frac{z \mu_e q^2 N_d^2 a^3}{\epsilon L} \quad [8.8]$$

y

$$V_P = \frac{q N_d a^2}{2\epsilon} \quad [8.9]$$

Tanto I_P como V_P son parámetros que dependen de las características del semiconductor utilizado así como de sus dimensiones geométricas. Así, un aumento de la sección transversal del semiconductor (esto es, de la semianchura a ó la profundidad z), o bien una disminución de la longitud L , hace que el valor de I_P sea más elevado. Del mismo modo, el voltaje V_P también crece con la semianchura del semiconductor. Este parámetro es conocido como *voltaje de estrangulamiento*, ya que su valor coincide aproximadamente (si no se tiene en cuenta el valor V_o) con el potencial en inverso que es necesario aplicar sobre la puerta para cerrar el canal cuando la corriente de drenador es cero (es decir $I_D=0$), según se puede demostrar fácilmente a partir de la ec. [8.4]. De esta misma ecuación se deduce además que si aplicamos a la puerta un voltaje determinado, V_G , la situación de estrangulamiento en el borde del drenador se pro-

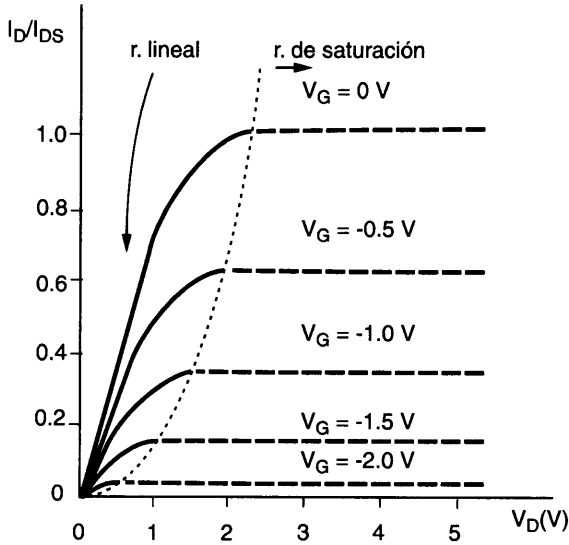


Fig. 8.6. Curvas características I_D - V_D normalizadas para un JFET de canal n , según la ec. [8.7], con $V_p = 3$ V.

duce cuando la tensión de drenador alcanza un valor, $V_{D,sat}$, tal que cumpla la ecuación:

$$a = \left[\frac{2\varepsilon}{q N_d} (V_o + V_{D,sat} - V_G) \right]^{1/2}$$

o bien, teniendo en cuenta la expresión [8.9]:

$$V_p = V_{D,sat} + V_o - V_G \quad [8.10]$$

la cual sirve para calcular $V_{D,sat}$.

En la fig. 8.6 se ha representado en trazo continuo las curvas teóricas I_D - V_D (normalizadas) correspondientes a la ec. [8.7] para un semiconductor típico. Las curvas representan la variación del cociente I_D/I_{DS} en función de V_D (siendo I_{DS} el valor de $I_{D,sat}$ correspondiente a $V_G = 0$) para voltajes inferiores al de saturación ($V_D < V_{D,sat}$). Obsérvese que para voltajes V_D pequeños las curvas características tienen un comportamiento cuasi-lineal, con una pendiente en el origen dada por el factor I_p / V_p . Es fácil demostrar que este cociente representa la

resistencia del canal completamente abierto. A medida que aumenta V_D las curvas se separan del comportamiento lineal debido a la influencia de los términos segundo y tercero de la ec. [8.7]. Para $V_D > V_{D,sat}$ la ec. [8.7] deja de ser válida y la corriente se considera constante, igual al valor de saturación, $I_{D,sat}$ (líneas de trazo discontinuo en la fig. 8.6). El valor de $I_{D,sat}$ se puede obtener a partir de la ec. [8.7] haciendo $V_D = V_{D,sat}$, con lo que resulta:

$$I_{D,sat} = I_P \left[\frac{V_P - 3(V_o - V_G)}{3V_P} + \frac{2}{3} \left(\frac{V_o - V_G}{V_P} \right)^{3/2} \right] \quad [8.11]$$

Utilizando los datos de la fig. 8.6 para las curvas características I_D - V_D de un transistor JFET se puede obtener la variación de I_D en función de V_G en la región de saturación. Resulta así la denominada *curva de transferencia* (fig. 8.7), la cual se obtiene a partir de la ordenada del punto de intersección de las curvas de la fig. 8.6 con una recta paralela al eje de ordenadas en un punto de abscisa V_D , contenido en la región de saturación. Es fácil demostrar que la curva de transferencia cumple aproximadamente la ecuación:

$$I_{D,sat} = I_{DS} \left(1 - \frac{V_G}{V_P} \right)^2 \quad \text{para } (V_G < V_P) \quad [8.12]$$

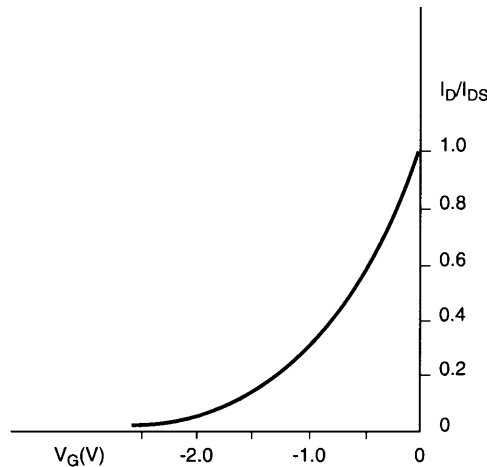


Fig. 8.7. Curva de transferencia para un JFET de canal n (I_D en función de V_G), obtenida a partir de las curvas de la figura anterior en la región de saturación.

donde I_{DS} es, según hemos visto, la corriente de saturación máxima, es decir, la corriente $I_{D,sat}$ para $V_G = 0$. Aunque esta ecuación parece muy distinta de la ec. [8.11], da sin embargo resultados muy aproximados y es mucho más sencilla de aplicar. A partir de la curva de transferencia se define un parámetro característico del transistor denominado *transconductancia*, g_m , el cual coincide con la pendiente de la curva de transferencia en el punto de operación, es decir:

$$g_m = \left. \frac{\Delta I_D}{\Delta V_G} \right|_{V_D} \quad [8.13]$$

Al estudiar los amplificadores con transistores (cap. IX) veremos que el factor de amplificación en voltaje de los JFET está directamente relacionado con el factor g_m .

8.3. CIRCUITO EQUIVALENTE DEL JFET PARA SEÑALES PEQUEÑAS (*)

Cuando se utiliza el transistor JFET como amplificador de señales alternas normalmente se emplea la *configuración de fuente común*, mostrada en la fig. 8.8a, polarizando el transistor con una tensión entre fuente y drenador, V_{DS} , suficientemente elevada para que el punto de operación se sitúe en la región de saturación (definido por una corriente I_D). La señal alterna que se desea amplificar, v_g , se aplica en el electrodo de puerta superpuesta a la tensión continua de polarización de puerta, V_{GS} . Esta señal de voltaje se traduce en variaciones de la corriente de drenador, i_d , que se superponen al valor continuo, I_D , y producen a su vez una variación, v_d , en la caída de tensión en la resistencia, R_L , introducida en el circuito de salida.

En la región de saturación del transistor se puede utilizar un análisis del comportamiento del JFET para señales pequeñas similar al del transistor bipolar funcionando en la región activa con la configuración de emisor común (apartado 6.5), haciendo la salvedad de que para el JFET la corriente en el circuito de entrada, esto es, la corriente de puerta, es prácticamente cero. Así pues, en este análisis sólo es preciso considerar el circuito de salida, en el cual la dependencia del valor instantáneo de la corriente de saturación, i_D , con las variables v_D y v_G puede escribirse como $i_D = i_D(v_D, v_G)$. Mediante el cálculo diferencial resulta para la componente alterna de i_D :

$$i_d = g_D v_d + g_m v_g \quad [8.14]$$

donde g_m es la transconductancia definida por ec. [8.13], y

$$g_D = \left. \frac{\partial I_D}{\partial V_D} \right|_{V_G} \quad [8.15]$$

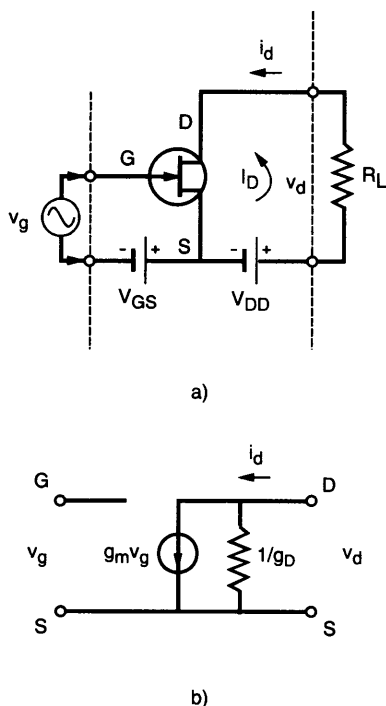


Fig.8.8. a) Circuito amplificador simple basado en un transistor JFET de canal n polarizado en la configuración de fuente común. b) Circuito equivalente de alterna para señales pequeñas del JFET funcionando en la región de saturación.

es la denominada *conductancia del canal*. En los transistores JFET esta magnitud suele tener un valor muy pequeño, ya que las curvas características en la región de saturación son prácticamente horizontales.

La ec. [8.14] sugiere el circuito equivalente de pequeña señal del JFET representado en la fig. 8.8b, en el cual la corriente en el circuito de salida, i_d , se obtiene mediante la suma de la corriente suministrada por el generador de corriente de valor $g_m v_g$ y la corriente a través de la resistencia $r=1/g_D$ debida a la señal v_d en el circuito de salida. Normalmente esta resistencia es muy elevada, varias decenas de kiloohmios, por lo que desde un punto de vista práctico suele eliminarse en el circuito de salida. Por otra parte, el circuito de entrada se ha dejado abierto, es decir sin conexión con el terminal de fuente o de drenador, para tener en cuenta **el hecho fundamental de que la resistencia de entrada en el terminal de puerta de los JFET es muy elevada**. Es ésta quizás una de las características más distintivas de la familia de transis-

tores de efecto campo. Nótese la similitud del circuito de la fig. 8.8b con el circuito equivalente del transistor bipolar representado en la fig. 6.18.

Cuando los JFET operan a muy alta frecuencia hay que tener en cuenta el tiempo de tránsito, t_r , de los portadores a través del canal. Se puede demostrar que t_r es proporcional a L^2 , por lo que la frecuencia de corte a partir de la cual los portadores ya no responden a las variaciones de la señal aplicada es inversamente proporcional a L^2 . Por tanto para mejorar el comportamiento del transistor a frecuencias altas es muy conveniente que la longitud del canal sea lo más pequeña posible.

8.4. EL TRANSISTOR DE UNION METAL-SEMICONDUCTOR (MESFET)

El MESFET fué propuesto por Mead en 1966, y aunque su funcionamiento es conceptualmente similar al JFET discutido más arriba, desde un punto de vista práctico puede operar a frecuencias bastante más altas, en la región de las microondas. A diferencia del JFET, el electrodo de puerta está formado por una unión metal-semiconductor (de ahí el nombre de MESFET) de tipo Schottky en lugar de una unión p-n, como se puede apreciar en el esquema

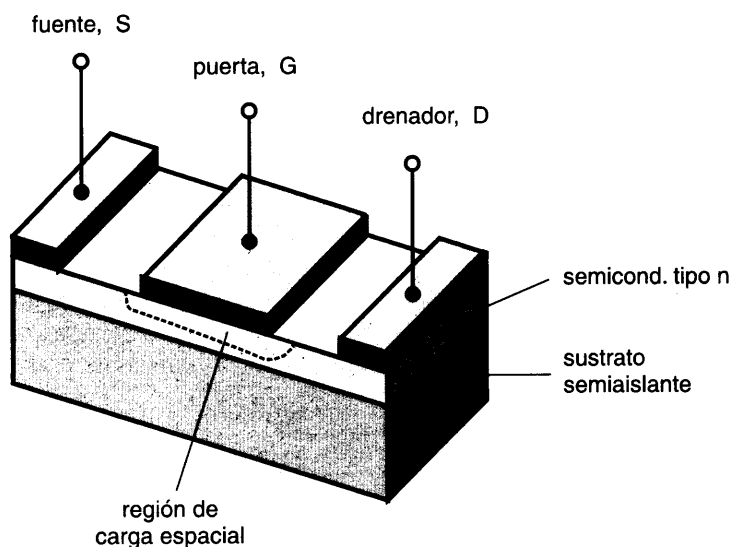


Fig. 8.9. Esquema de la estructura de un transistor tipo MESFET. El contacto de puerta está formado por una unión metal-semiconductor, tipo Schottky.

de la fig. 8.9. La estructura se completa con otros dos electrodos metálicos depositados sobre la superficie del semiconductor, en sus extremos, formando un contacto óhmico con el semiconductor. Uno de estos electrodos actúa de *fuentes o surtidor* (S) y el otro de *drenador o sumidero* (D).

Los MESFET son contruidos en casi su totalidad a partir de arseniuro de galio en lugar de silicio. Las ventajas de la utilización del arseniuro de galio son varias: i) En el rango de operación útil, los electrones del arseniuro de galio presentan una movilidad de $0.8 \text{ m}^2\text{V}^{-1}\text{s}^{-1}$, es decir unas cinco veces superior a la del silicio. ii) La velocidad de arrastre máxima de los electrones por un campo eléctrico es en el GaAs alrededor del doble de la de los electrones en el Si (véase la fig. 8.10). iii) Los MESFET se fabrican depositando una capa epitaxial de GaAs dopada convenientemente sobre un sustrato de GaAs con propiedades semiaislantes ($\rho \approx 10^8 \text{ ohm cm}$). De este modo, las capacidades parásitas entre el sustrato y los contactos metálicos de los electrodos son muy bajas. iv) La posibilidad, en los dispositivos de puertas muy cortas, de que los electrones alcancen velocidades muy elevadas, lo que resulta en un tiempo de tránsito muy pequeño. Todo ello hace posible construir hoy día amplificadores que pueden operar hasta frecuencias de 60 GHz, así como circuitos digitales de muy alta velocidad. Debido a todas estas características el GaAs está sustituyendo ventajosamente al Si en algunas aplicaciones especiales, a pesar de que la tecnología de circuitos integrados basada en el GaAs está todavía mucho menos desarrollada que la del Si.

8.5. CURVAS CARACTERISTICAS INTENSIDAD-VOLTAJE DEL MESFET

El funcionamiento del MESFET es muy similar al del JFET estudiado en los apartados anteriores. De hecho la capa epitaxial semiconductor actúa como un canal efectivo para los portadores, ya que está limitado en su parte inferior por el sustrato semiaislante. Si el canal es de tipo n, el electrodo de puerta se polariza negativamente, con lo cual se forma una región de carga espacial en el semiconductor vacía de portadores, que controla la anchura efectiva del canal. Asimismo, en el modo normal de operación, la fuente se conecta a tierra y el drenador a un potencial positivo. De esta forma la fuente inyecta electrones hacia el sumidero a través del canal formado por el semiconductor.

La fig. 8.11 muestra la variación de la intensidad I_D en función del voltaje V_D aplicado al drenador de un MESFET típico. Consideremos primero que la tensión de puerta se mantiene a potencial cero (esto es $V_G = 0$). Para pequeños valores de V_D el canal semiconductor actúa como una resistencia pura, y por tanto la relación $I_D - V_D$ es de tipo lineal. Sin embargo, cuando se aplican voltajes V_D más elevados, el canal semiconductor se hace cada vez más positivo respecto de la puerta por lo que la región de carga espacial se hace progresivamente más ancha. Esto implica una disminución de la anchura del canal (sobre todo en la zona próxi-

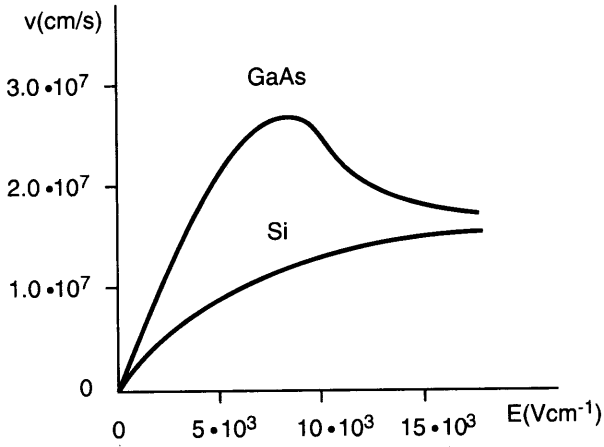


Fig. 8.10. Velocidad de arrastre de los electrones en función del campo eléctrico, para el GaAs y el Si.

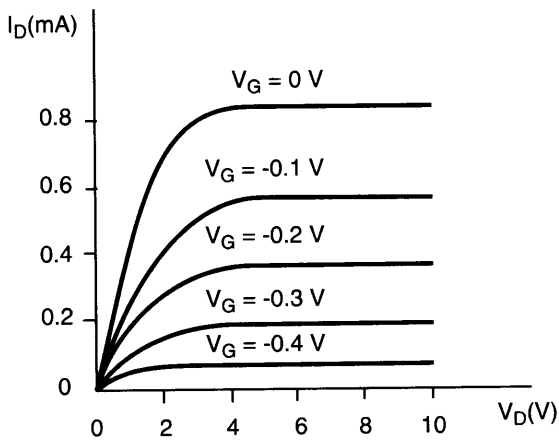


Fig. 8.11. Curvas características I_D - V_D de un MESFET (con un semiconductor tipo n), de canal "normalmente abierto".

ma al drenador) y un aumento de su resistencia, lo cual da lugar a una disminución de la pendiente de la curva $I_D - V_D$. Cuando V_D toma valores superiores a una cierta tensión crítica o umbral, necesaria para el estrangulamiento del canal, la corriente se estabiliza por un efecto similar al JFET (fig. 8.11).

En el caso de los MESFET de GaAs, este efecto de saturación de la corriente es debido en gran parte a la limitación de la velocidad de los electrones en la región del canal próxima al drenador. Debido al diseño de los MESFET, con una longitud de canal generalmente muy pequeña, el campo eléctrico en la región del sumidero puede alcanzar valores muy elevados cuando se aplican tensiones de drenador suficientemente altas. La velocidad de los electrones se sitúa entonces a la derecha del máximo en la curva correspondiente al GaAs en la fig. 8.10,

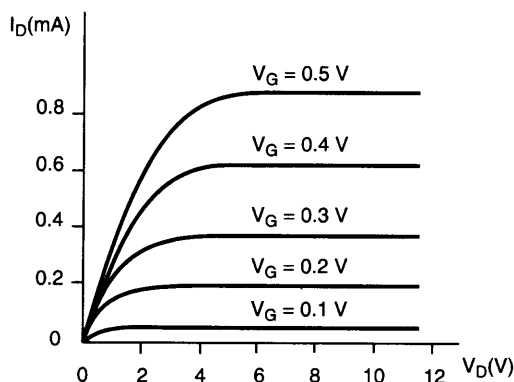


Fig. 8.12. Curvas características $I_D - V_D$ de un MESFET similar al anterior, de canal "normalmente cerrado".

alcanzando un valor de saturación. En esta región de campos eléctricos elevados, un aumento del campo eléctrico se compensa por una disminución de la movilidad de los electrones, con lo que la corriente a través del canal se mantiene constante.

Evidentemente, si el potencial de puerta, V_G , se hace negativo la región de carga espacial se ensancha y el canal semiconductor se hace más estrecho. Por esta razón, la corriente I_D a través del canal será menor cuanto más negativo sea V_G . Se puede obtener una relación cuantitativa para la característica $I_D - V_D$ siguiendo el mismo cálculo que se efectuó en el apartado anterior para el JFET. De hecho el resultado de la ec. [8.7] es también válido para el MESFET, siendo V_p el voltaje crítico o umbral a partir del cual la corriente se estabiliza por efecto del estrangulamiento del canal.

Cuando se pretende una velocidad de respuesta elevada a menudo se recurre a dispositivos MESFET diseñados de tal manera que el canal se encuentra cerrado cuando la tensión aplicada es cero. Para ello, basta que el espesor de la capa epitaxial depositada sobre el sustrato semiaislante y que constituye el canal sea suficientemente fina. Cuando este espesor es menor que el de la carga espacial de la unión metal-semiconductor en equilibrio ($V_G = 0$) el canal está bloqueado, y el transistor se dice que es del tipo "normalmente cerrado" (en contraste a los anteriormente descritos que son de canal "normalmente abierto"). Sin embargo, si se aplican voltajes positivos con un valor pequeño a la puerta la región de carga espacial se reduce dejando libre un pequeño canal para la corriente que pasa desde la fuente al drenador. La tensión positiva aplicada a la puerta nunca puede ser superior a unas pocas décimas de voltio, ya que en otro caso se perdería el carácter bloqueante del contacto y una cierta fracción de la corriente a través del canal se perdería por el propio contacto de puerta. Las características I_D - V_D de estos dispositivos son muy similares a las de canal abierto, con la única diferencia de que la corriente aumenta al aumentar V_G , según se observa en la fig. 8.12.

8.6. TRANSISTORES METAL-OXIDO-SEMICONDUCTOR DE EFECTO CAMPO (MOSFET)

En el capítulo anterior se ha considerado la estructura MOS como un dispositivo de dos terminales, lo cual permite estudiar la estructura de bandas y la distribución de carga en el interior del semiconductor cuando se aplica una tensión en los extremos de la estructura. Estos estudios son necesarios para el entendimiento, e incluso la previsión, del funcionamiento de los transistores MOSFET, o abreviadamente MOS. Los transistores MOS encuentran en la actualidad amplia aplicación en las puertas lógicas utilizadas en electrónica digital y en las memorias semiconductoras.

8.6.1. Estructura básica del MOSFET

En la fig. 8.13a se muestra la estructura básica de un transistor MOS de silicio de tipo p, desarrollado también según la tecnología planar. Básicamente consiste en una estructura MOS en la cual el electrodo metálico superior, G, depositado sobre la capa aislante actúa como terminal de puerta del transistor. Existen además dos regiones pequeñas de la superficie dopadas fuertemente con impurezas donadoras, es decir de tipo n^+ , situadas a cada lado de la puerta. Sobre cada una de estas regiones o islas de tipo n^+ se deposita asimismo un electrodo metálico, formando el contacto de fuente o surtidor, S, y el de sumidero o drenador, D, del transistor. Finalmente, al igual que en una estructura MOS simple, sobre la superficie inferior del dispositivo se deposita una capa metálica que se mantiene conectada a tierra. En la fig. 8.13b se muestra el símbolo del transistor MOS de canal n (como el de fig. 8.13a) en el cual la flecha indica el sentido convencional de la corriente (de drenador a fuente) en el modo normal

de operación del transistor. En el MOSFET de canal p, la corriente tiene sentido opuesto y la flecha del dibujo lleva la dirección invertida.

8.6.2. Descripción cualitativa del funcionamiento del MOSFET

Consideremos primero el caso de que el voltaje aplicado a la puerta es cero, es decir, $V_G = 0$. Las dos regiones o islas de tipo n^+ de fuente y drenador forman con el resto del semiconductor de tipo p sendas uniones p- n^+ conectadas en oposición, por lo que prácticamente no existe paso de corriente entre los electrodos de fuente y sumidero, cualquiera que sea el signo de la tensión aplicada entre ellos.

Supongamos ahora que aplicamos un voltaje positivo suficientemente alto a la puerta para tener la condición de inversión fuerte en la interfase del semiconductor con el óxido (véase apartado 7.1). Esto quiere decir que $V_G \geq V_T$, siendo V_T el voltaje umbral de la estructura MOS. Los portadores minoritarios, electrones en este caso, dan lugar a la formación de un canal conductor de tipo n en la superficie del semiconductor entre la fuente y el drenador con una conductancia mayor cuanto más alto sea el voltaje aplicado en la puerta. Como ya

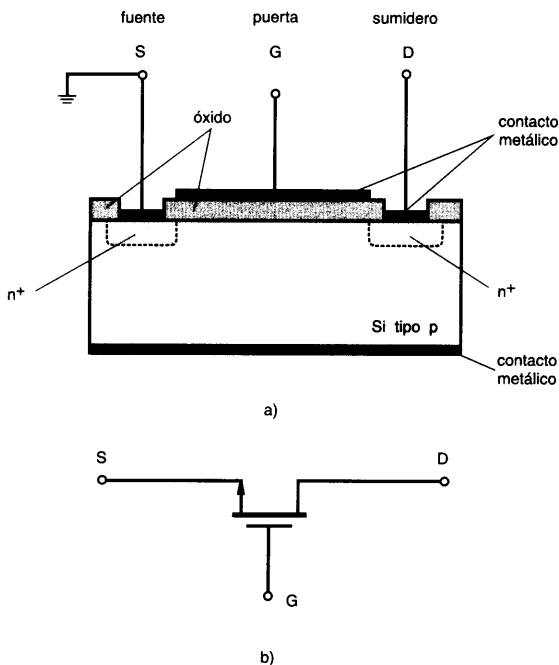


Fig. 8.13. a) Estructura típica de un MOSFET de canal n. b) Símbolo del MOSFET de canal n utilizado en circuitos.

vimos en el capítulo anterior, este canal está limitado en su parte superior por la capa aislante de SiO_2 y en la parte inferior por la región de carga espacial que se forma en el semiconductor bordeando la puerta y también alrededor de las islas n^+ de los contactos de fuente y sumidero. En estas circunstancias si se aplica un voltaje positivo de valor pequeño al drenador (véase la fig. 8.14a), los electrones fluyen desde la fuente al drenador a lo largo del canal que actúa ahora como si fuera una resistencia de valor bajo (nótese que en el canal, al estar invertido, la conducción tiene el mismo carácter que en la fuente y el sumidero). Se obtiene así, en esta región de voltajes de drenador bajos, una variación lineal entre la corriente I_D y el voltaje aplicado, V_D .

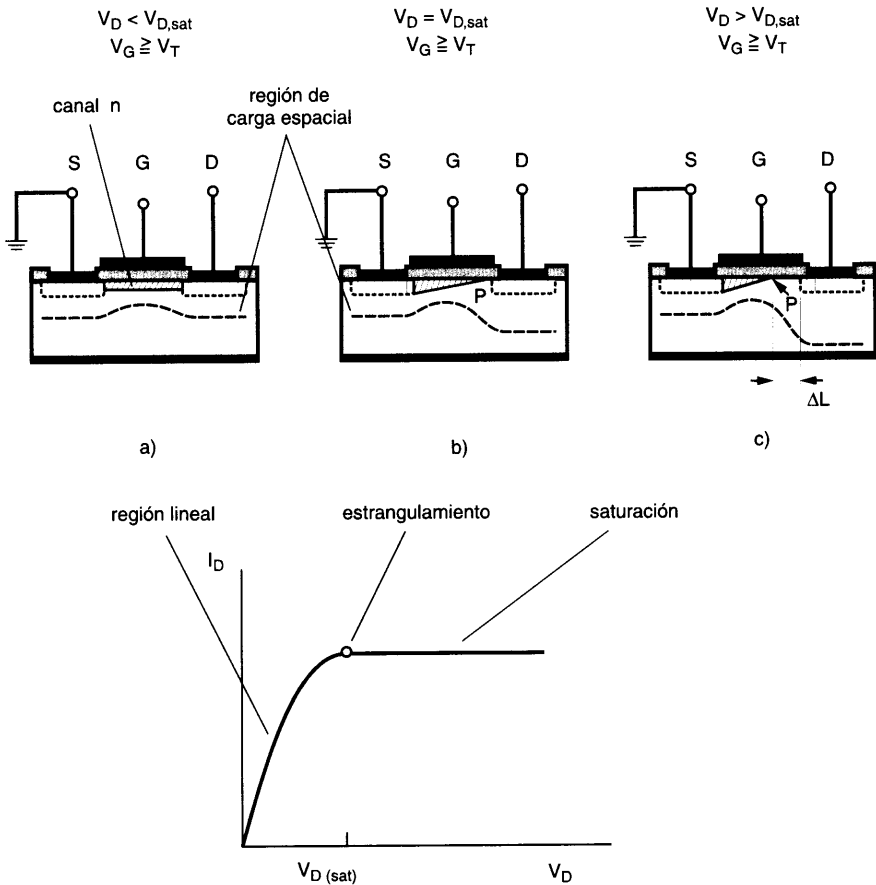


Fig. 8.14. Esquema de la variación de la anchura del canal al aplicar voltajes de drenador crecientes en un MOSFET de canal n. En la parte inferior se muestra la correspondiente curva característica I_D - V_D .

Al aumentar V_D , manteniendo $V_D < V_G$, tanto la región de carga espacial que rodea al drenador como aquella que existe a lo largo del canal se ensancha, ya que el contacto inferior del dispositivo está a tierra. Debido a la progresiva caída de tensión desde la fuente al drenador, el ensanchamiento es tanto mayor cuanto mas próximos nos hallemos del drenador. Por otra parte, esta caída de tensión a lo largo del canal semiconductor hace que la diferencia de potencial efectiva entre la puerta y el semiconductor sea cada vez más pequeña a medida que se avanza hacia el drenador. **Todo ello da lugar a una disminución del número de electrones que están presentes en la capa de inversión próxima al drenador, lo que equivale a su vez a una reducción, también progresiva, de la anchura del canal.** Evidentemente, esta reducción es más acusada en la zona del drenador. El efecto global es una disminución de la pendiente en la curva de variación de I_D en función de V_D .

Cuando se alcanza un voltaje, tal que la anchura del canal se reduce a cero en el drenador (fig. 8.14b), se dice que ha ocurrido el estrangulamiento del canal (punto P en fig. 8.14b). Esto ocurrirá para un voltaje denominado *voltaje de saturación*, $V_{D,sat}$, el cual ha de cumplir la relación $V_{D,sat} = V_G - V_T$. Para voltajes más elevados, la región del canal estrangulada, ΔL , aumentará de longitud en la dirección de la fuente (fig. 8.14c) y la corriente se mantendrá esencialmente constante, ya que el voltaje en el nuevo punto P de estrangulamiento se mantiene prácticamente igual a $V_{D,sat}$. De hecho, el mecanismo de limitación de la corriente entre el punto P y la región de agotamiento del drenador es muy similar a la de un transistor JFET (véase sec. 8.1.2), de ahí que las características I_D - V_D sean también similares.

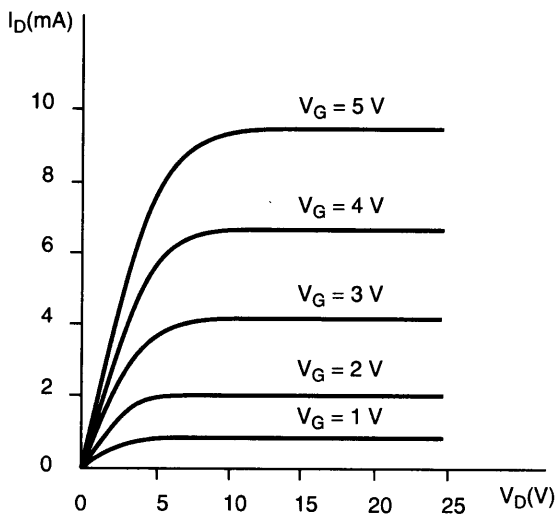


Fig.8.15. Curvas I_D - V_D típicas de un MOSFET de canal n.

En la fig. 8.15 se presentan las curvas experimentales I_D - V_D obtenidas al variar la tensión de puerta, V_D , en un transistor MOS de canal n. Obsérvese en esta figura que, para un voltaje V_D dado, al hacer más positiva la tensión V_G aplicada a la puerta la corriente I_D se hace más elevada, ya que el canal es cada vez más ancho y por tanto el número de portadores más elevado. Hay que notar también que para conseguir que el canal sea conductor es preciso que V_G tenga un valor por lo menos igual al voltaje umbral (alrededor de un voltio) para la formación en la superficie de una capa con fuerte inversión. Evidentemente, en los transistores MOS de silicio de tipo n el canal conductor es de tipo p y el potencial de puerta ha de ser negativo. Asimismo, la tensión V_D debe ser negativa en este caso. Por lo demás el comportamiento es exactamente igual al transistor MOS de canal n discutido más arriba.

8.7. CALCULO DE LAS CARACTERISTICAS INTENSIDAD-VOLTAJE DEL MOSFET(*)

Con objeto de obtener una relación sencilla entre la corriente I_D y el voltaje V_D para un transistor MOSFET de canal n consideraremos el caso ideal, es decir, cuando no hay diferencia entre la función de trabajo del metal de puerta y la del semiconductor, las cargas eléctricas en el óxido son nulas y no existen estados superficiales. Supondremos además que inicialmente aplicamos en la puerta una tensión V_G superior a la umbral, esto es, $V_G > V_T$. Esto quiere decir que, incluso sin tensión aplicada en el drenador, el canal se encuentra en condiciones de inversión fuerte.

Análogamente al caso del JFET estudiado anteriormente, consideraremos el caso en el que el transistor se polariza en la región cuasi-lineal de las curvas I_D - V_D . Esto implica que la tensión aplicada al drenador, V_D , (con la fuente a potencial cero) es menor que la tensión de saturación, $V_{D,sat}$, es decir $V_D < V_{D,sat}$. Siguiendo el modelo desarrollado para el JFET podemos suponer también que la anchura del canal es variable, disminuyendo en la dirección del drenador (en realidad lo que varía es la concentración de carga presente en cada punto del canal). Consecuentemente, la resistividad del canal debe aumentar al pasar de la fuente al drenador. Sabiendo que la corriente en cualquier punto del canal, I_D , debe ser independiente de la posición, podemos poner:

$$dV = I_D dR \quad [8.16]$$

siendo dR la resistencia de un elemento dx en una posición x a lo largo del canal, con una resistividad ρ variable. El valor de dR vendrá dado por:

$$dR = \rho(x) \frac{dx}{S} = \frac{l}{q \mu_e n(x)} \frac{dx}{z y(x)} \quad [8.17]$$

donde $S = zy(x)$ representa el área de la sección transversal del canal de profundidad z y altura $y(x)$ variable (fig. 8.16). El producto $qn(x)y(x)$ del denominador corresponde a la densidad de carga de inversión contenida en el canal en la posición x , $Q'_{inv}(x)$. Así pues, a partir de las ecs. [8.16] y [8.17] podremos escribir:

$$I_D dx = z \mu_e Q'_{inv}(x) dV \quad [8.18]$$

La carga por unidad de área en la zona de inversión del semiconductor, $Q'_{inv}(x)$, se puede expresar como (véase sec. 7.3.1):

$$Q'_{inv}(x) = Q'_s(x) - Q'_{agot}(x) \quad [8.19]$$

donde $Q'_s(x)$ es la carga total en el punto de coordenada x de la interfase del semiconductor con el aislante y $Q'_{agot}(x)$ es la carga comprendida en la zona de agotamiento (ambas por unidad de superficie). Ahora bien, en la situación de inversión Q'_s está directamente relacionada con la capacidad específica de la estructura MOS, C'_{ox} , a través de la ecuación: $Q'_s = -C'_{ox}V_{ox}$. En esta ecuación V_{ox} es la parte del potencial de puerta que cae en el óxido. Teniendo en cuenta además la ec. [7.20], la ecuación anterior queda:

$$Q'_{inv}(x) = -C'_{ox}[V_G - \psi_s(x)] - Q'_{agot} \quad [8.20]$$

siendo V_G el voltaje aplicado a la puerta y $\psi_s(x)$ el potencial de superficie correspondiente al punto de coordenada x . Por otra parte, en una estructura MOS con semiconductor tipo p, dopado con una concentración de impurezas N_a , la carga de agotamiento está relacionada con ψ_s a través de la ec. [7.16], por lo que la expresión anterior resulta:

$$Q'_{inv}(x) = -C'_{ox}[V_G - \psi_s(x)] + [2qeN_a\psi_s(x)]^{1/2} \quad [8.21]$$

Si denominamos $V(x)$ a la caída de voltaje entre la fuente del MOSFET y un punto de abscisa x a lo largo del canal, podemos escribir para el potencial de superficie del semiconductor cuando está en inversión fuerte: $\psi_s(x) \approx 2\psi_i + V(x)$, con ψ_i definido a través de la ec. [7.7]. Esta expresión de $\psi_s(x)$ sustituida en [8.21] nos da para $Q'_{inv}(x)$:

$$Q'_{inv}(x) = -C'_{ox}[V_G - 2\psi_i - V(x)] + \{2qeN_a[2\psi_i + V(x)]\}^{1/2} \quad [8.22]$$

En la fig. 8.16 se da un esquema cualitativo de la variación de la carga de inversión y del potencial a lo largo del canal en un transistor MOSFET de canal n polarizado en la región lineal.

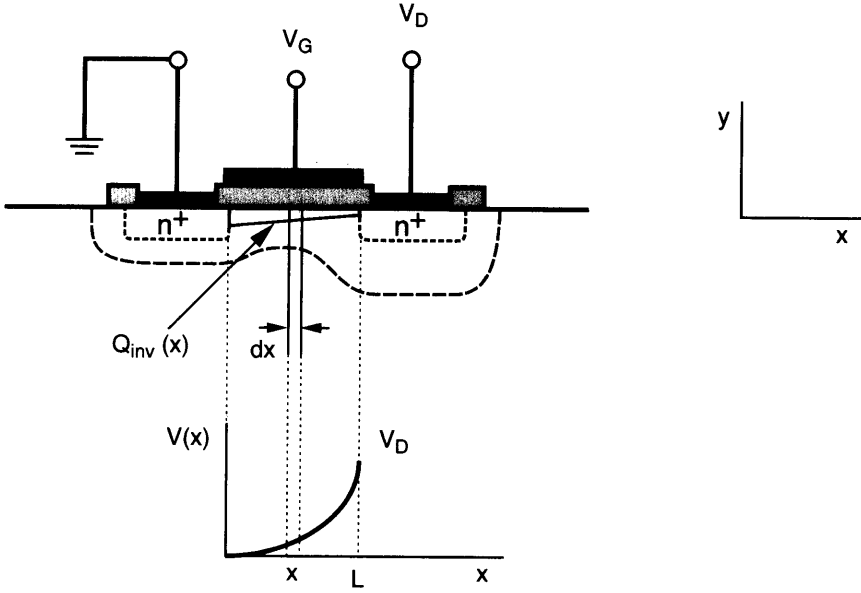


Fig.8.16. Esquema de la variación de la carga de inversión, $Q'_{inv}(x)$, y de la tensión, $V(x)$, a lo largo del canal en un MOSFET, tipo n, polarizado en la región lineal ($V_D < V_{D,sat}$).

Sustituyendo ahora en la ec. [8.18] el valor de $Q'_{inv}(x)$ dado por ec. [8.22] e integrando ambos miembros de la ecuación anterior entre los límites $x=0$ y $x=L$, para $V=0$ y $V=V_D$, respectivamente, resulta finalmente:

$$I_D = \frac{z}{L} \mu_e \left\{ C'_{ox} \left(V_G - 2\psi_i - \frac{V_D}{2} \right) V_D - \frac{2}{3} (2q\epsilon N_a)^{1/2} [(2\psi_i + V_D)^{3/2} - (2\psi_i)^{3/2}] \right\} \quad [8.23]$$

En el anterior análisis se ha considerado la movilidad constante con un valor igual al valor medio de la movilidad de los electrones dentro del canal. Debido a que la anchura del canal es muy pequeña, los electrones de la capa de inversión interaccionan fuertemente con la superficie del óxido, disminuyendo su velocidad. Por esta razón la movilidad, μ_e , en el canal es más baja que en el interior del semiconductor, y además puede incluso depender de la anchura del canal.

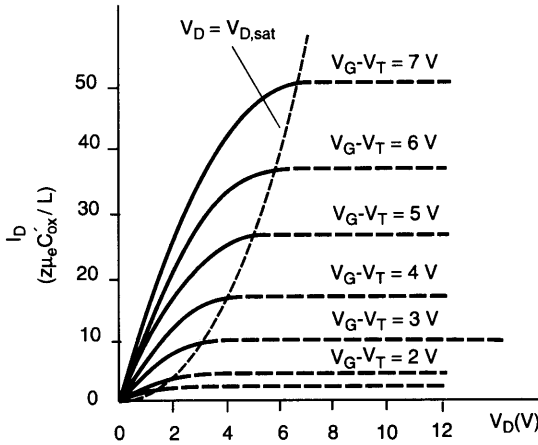


Fig. 8.17. Curvas características I_D - V_D (normalizadas) para un MOSFET de canal n, según la ec. [8.23].

En la fig. 8.17 se ha representado en trazo continuo las curvas características normalizadas intensidad-voltaje de un MOSFET según la ec. [8.23]. Según se observa, para pequeños valores de V_D la dependencia es de tipo lineal, mientras que para voltajes más elevados la pendiente disminuye hasta el valor cero, correspondiente al punto de abscisa $V_D = V_{D,sat}$. A pesar de las aproximaciones empleadas en el modelo, también en este caso se obtiene un buen acuerdo entre las curvas teóricas y las obtenidas experimentalmente (fig. 8.15)

En la región lineal, para voltajes V_D mucho menores que $(V_G - V_T)$ la ec.[8.23] puede aproximarse por:

$$I_D \approx \frac{z}{L} \mu_e C'_{ox} (V_G - V_T) V_D \quad [8.24]$$

En esta región la conductancia del canal, g_D , vendrá dada de forma aproximada por:

$$g_D = \left. \frac{\partial I_D}{\partial V_D} \right|_{V_G} \approx \frac{z}{L} \mu_e C'_{ox} (V_G - V_T) \quad [8.25]$$

y la transconductancia, g_m :

$$g_m = \left. \frac{\partial I_D}{\partial V_G} \right|_{V_D} \approx \frac{z}{L} \mu_e C'_{ox} V_D \quad [8.26]$$

Al igual que en el caso del JFET, para $V_D = V_{D,sat}$ ocurre el estrangulamiento del canal en el drenador. Por encima de este voltaje, la corriente toma un valor constante, $I_{D,sat}$ (líneas de trazo discontinuo en la fig. 8.17). El valor de la corriente de saturación, $I_{D,sat}$, puede hallarse obteniendo primero el valor de $V_{D,sat}$ a partir de la ec. [8.22] imponiendo la condición que para $x=L$ se cumple que $Q'_{inv} = 0$ y sustituyendo el valor así obtenido en la ec. [8.23] (véase problema 8.9). Se obtiene de este modo para $I_{D,sat}$:

$$I_{D,sat} \approx \frac{z \mu_e \epsilon_{ox}}{2dL} (V_G - V_T)^2 \quad [8.27]$$

donde d y ϵ_{ox} son, respectivamente, el espesor y la constante dieléctrica del óxido y V_T es el voltaje umbral dado por la ec. [7.24] del capítulo anterior. Evidentemente, en la región de saturación la conductancia del canal es prácticamente nula, mientras que la transconductancia g_m viene dada, a partir de ec. [8.26], por:

$$g_m = \left. \frac{\partial I_D}{\partial V_G} \right|_{V_D} \approx \frac{z \mu_e \epsilon_{ox}}{dL} (V_G - V_T) \quad [8.28]$$

siendo una función que depende únicamente de la tensión aplicada en la puerta, V_G .

8.8. CIRCUITO EQUIVALENTE DEL MOSFET PARA SEÑALES PEQUEÑAS (*)

En los circuitos amplificadores de señales alternas basados en un MOSFET, el transistor se polariza en la región de saturación utilizando la configuración de fuente común. La señal alterna, v_g , que se pretende amplificar se introduce, al igual que en los otros tipos de transistores estudiados en este capítulo, superpuesta a la tensión de polarización de puerta, V_G . En estos circuitos, las variaciones producidas en la tensión de puerta originan a su vez una variación en la corriente de drenador, i_d , que se superpone a la corriente continua del drenador, I_D . Estas variaciones de corriente provocan también una señal variable, v_d , en la resistencia R_L introducida en el circuito de salida.

En la región de saturación se puede realizar un análisis del comportamiento del MOSFET similar al del transistor bipolar funcionando para la región activa. Las características del

circuito equivalente resultante son muy similares a las obtenidas para el transistor JFET. La fig. 8.18a muestra el circuito equivalente de un MOSFET operando a baja frecuencia. Igual que en el JFET, la puerta G se encuentra separada de los electrodos de la fuente y el sumidero debido a la resistencia prácticamente infinita de la capa de óxido. En este sentido, el MOSFET es el transistor que presenta mejores características en lo que se refiere a la resistencia de entrada, ya que al ser casi infinita, el dispositivo no consume corriente de la fuente de alimentación que proporciona la señal v_g . Nótese en el circuito equivalente que la señal de salida se obtiene a través de un generador de corriente de valor $g_m v_g$. Hay que señalar que, en lo que se refiere al circuito de salida, el MOSFET, al igual que otros transistores, se comporta como una fuente de corriente constante. Este hecho es debido a la horizontalidad de las curvas I_D - V_D en la región de saturación.

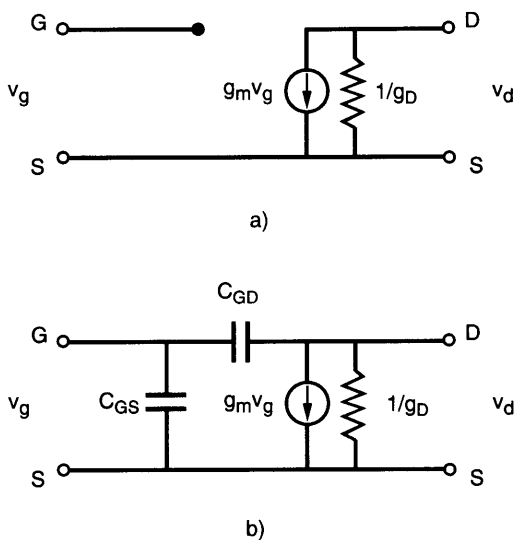
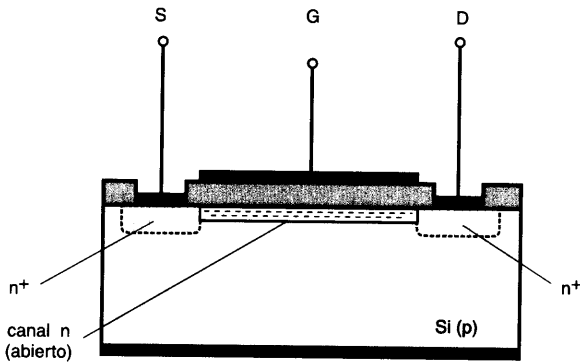


Fig.8.18. a) Circuito amplificador simple de señales pequeñas para un MOSFET, de canal n, polarizado en la región de saturación. b) Circuito equivalente del transistor para baja frecuencia.

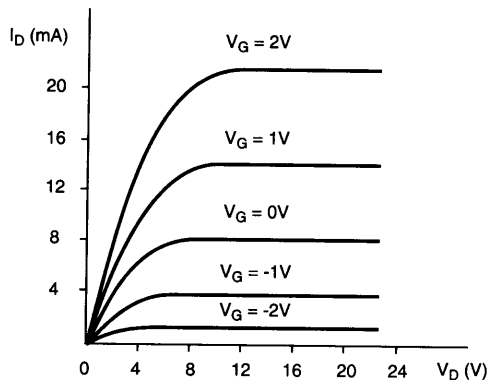
Cuando se trabaja con señales de alta frecuencia es preciso tener en cuenta las capacidades parásitas entre el electrodo de puerta y los de fuente y sumidero (C_{GS} y C_{GD} en la fig. 8.18b). Estas capacidades parásitas pueden tener bastante importancia en el transistor MOSFET, ya que alteran profundamente el tiempo de respuesta del transistor. Se puede demostrar que el tiempo de respuesta es proporcional al cuadrado de la longitud L del canal, por lo que cuando se desea trabajar a frecuencias altas es preciso reducir la longitud del canal al máximo posible.

8.9. OTROS TIPOS DE MOSFET (*)

En las secciones anteriores hemos visto que la conductividad del canal en un MOSFET aumenta sustancialmente al formarse una capa de inversión mediante la aplicación de un voltaje superior al umbral. A este tipo de transistores, bien sean de canal n ó p se les denomina *MOSFET de enriquecimiento* y también se les conoce como transistores de canal "normalmente cerrado".



a)



b)

Fig. 8.19. a) Estructura de un MOSFET de canal n de agotamiento. b) Curvas características I_D - V_D para un MOSFET típico de canal n de agotamiento.

Se puede conseguir que el transistor funcione con el canal «normalmente abierto», si la región del canal se dopa con impurezas de tipo opuesto al resto del semiconductor, formando así una capa conductora entre las islas que forman las regiones de fuente y sumidero. A este tipo de transistores se les denomina *MOSFET de agotamiento*, ya que como veremos después el transistor puede funcionar en la región de agotamiento del canal. En la fig. 8.19a se ha representado un MOSFET de agotamiento con canal n. Obsérvese que el canal conductor, al igual que las regiones de puerta y drenador es de tipo n. La aplicación de voltajes negativos a la puerta da lugar en este caso a una disminución del número de portadores en el canal (región de agotamiento) por lo que la corriente I_D disminuye para un voltaje V_D fijo, pudiendo incluso hacerse cero si V_G es suficientemente negativo. Al contrario, si el voltaje de puerta es positivo, con un valor superior al umbral, el canal se enriquece con los portadores minoritarios de inversión y la corriente aumenta. Lógicamente las características I_D - V_D de un MOSFET de agotamiento son similares a las del MOSFET de enriquecimiento con la diferencia de que la tensión de puerta admite a la vez valores positivos y negativos (fig. 8.19b).

8.10. ASPECTOS TECNOLOGICOS DEL MOSFET

Igual que ocurre en el JFET, el límite de la frecuencia superior de utilización del transistor está relacionado con el tiempo de tránsito de los electrones de la fuente al sumidero, por lo que es conveniente que la longitud L del canal sea la menor posible, siendo el límite actual de L alrededor de media micra. De los dos tipos de MOSFET de silicio, de canal n o de canal p, los primeros son preferidos por la mayor movilidad de los electrones en relación a los huecos.

El voltaje umbral para la estructura MOS ideal, dado por la ecuación [7.24], es alrededor de 1V para los MOSFET de silicio de canal n y -1V para los de canal p. Ahora bien, si tenemos en cuenta que la diferencia entre las funciones de trabajo del metal y el semiconductor es $\phi_{ms} \approx -1V$ y que el cociente Q_f/C_{ox} (Q_f es la carga fija en el óxido) resulta también del orden de -1V, se tendrá que el voltaje umbral real es alrededor de -1V y -3V para los MOSFET de canal n y p, respectivamente. Quizás uno de los mayores avances en la disminución de V_T se consiguió al utilizar el propio silicio en forma policristalina (polisilicio) como contacto de puerta, en lugar del aluminio, ya que de esta forma la diferencia de funciones de trabajo es cero. Para hacer más conductor el silicio policristalino se dopa fuertemente con fósforo durante la deposición, quedando el material como un semiconductor degenerado, es decir con una conductividad elevada.

Para disminuir aún más el valor de V_T se siguen diversos procedimientos. El primero consiste en utilizar para el sustrato de silicio obleas cortadas según la cara [100] de la superficie en lugar de la cara [111]. De este modo, se consigue así que el valor de Q_f disminuya notablemente (véase apartado 7.4). Otro procedimiento consiste en aumentar la capacidad

C_{ox} utilizando una primera capa de SiO_2 para conservar las buenas propiedades de la interfase Si- SiO_2 seguido de una capa de nitruro de silicio (Si_3N_4) que tiene una constante dieléctrica aproximadamente el doble de la del SiO_2 . Otro método muy utilizado para reducir el voltaje umbral consiste en dopar ligeramente la región del canal con impurezas de signo opuesto mediante implantación iónica. Para ello iones de tipo donador, fósforo, p.e., son acelerados hasta unos 300 KeV y con ellos se bombardea el sustrato de tipo p, siendo más fácil entonces formar el canal n. Mediante una dosificación adecuada de los iones donadores, con esta técnica se puede formar en un mismo proceso las regiones o islas de la fuente y el sumidero (de carácter n^+). Incluso, en el mismo sustrato de tipo p se pueden fabricar también otros dispositivos MOSFET de agotamiento simplemente aumentando la dosis de fósforo en la región correspondiente al canal de estos dispositivos. Por último mencionaremos que también se puede disminuir V_T haciendo el aislante de la puerta lo más fino que sea posible. En este sentido, actualmente se han preparado transistores MOS con espesor del óxido aislante de tan sólo unos 100 Å. Para conseguir que un óxido de estas características mantenga sus propiedades aislantes es preciso preparar la capa de óxido con un grado de perfección elevado. Los problemas asociados con la técnica de obtención del SiO_2 vienen descritos en el capítulo XIII.

CUESTIONES Y PROBLEMAS

- 8.1 Señalar las características más destacadas de los transistores de efecto campo, comparándolas con las de los transistores bipolares.
- 8.2 Si un transistor de efecto campo se encuentra en condiciones de estrangulamiento del canal, es decir, con $V_D = V_{D,sat}$, ¿cómo es posible que haya paso de corriente a través de canal?
- 8.3 Hacer un esquema de las bandas de energía a lo largo de una línea vertical en la sección transversal del JFET especificada en la fig. 8.5.
- 8.4 Sea un transistor JFET de silicio de canal n, con una semianchura $a = 2,5 \mu m$ y un dopaje $N_d = 10^{15} cm^{-3}$. Se pide calcular: a) El voltaje de estrangulamiento y b) la semianchura del canal cuando $I_D = 0$ y $V_G = 2V$. Comparar el resultado con la semianchura original.
- 8.5 Hallar la resistencia dinámica r_d correspondiente a las curvas $I_D - V_D$ de un JFET en el origen, y demostrar que viene dada por:

$$r_d = \frac{L}{2azqN_d\mu_e [1 - (V_G / V_p)^{1/2}]}$$

Asimismo, para el transistor de curvas características como las de la fig. 8.4, hallar de forma aproximada el valor de r_d para $V_G = 0$.

- 8.6** Sea un transistor de silicio de efecto campo de unión de dimensiones: $a=0.85 \mu\text{m}$, $L=25 \mu\text{m}$ y $z=120 \mu\text{m}$ para el cual el dopaje del canal y de la puerta es $N_d=10^{16} \text{cm}^{-3}$ y $N_a=10^{19} \text{cm}^{-3}$, respectivamente. Calcular el voltaje de estrangulamiento y la corriente correspondiente para el caso $V_G=0$. (Tómese $\mu_e=1350 \text{cm}^2\text{V}^{-1}\text{s}^{-1}$ y $\epsilon=11.9$).
- 8.7** Dibujar la curva de transferencia de un JFET con $I_{DS}=12 \text{mA}$ y $V_p=-5\text{V}$. Demostrar que la tangente a esta curva en el punto de corte con el eje de ordenadas corta al eje de abscisas en el punto $V_p/2$.
- 8.8** En un transistor de GaAs tipo MESFET de canal n con $N_d=10^{17} \text{cm}^{-3}$ la altura de la barrera metal-semiconductor en el electrodo de puerta es de 0.9eV , y las dimensiones del canal son: $a=0.2 \mu\text{m}$, $L=1 \mu\text{m}$ y $z=10 \mu\text{m}$. a) Averiguar si se trata de un dispositivo con canal normalmente abierto o cerrado, b) calcular el voltaje umbral y la corriente de saturación para $V_G=0$.
- 8.9** A partir de la expresión [8.23] demostrar que para un MOSFET de canal n el voltaje de saturación viene dado por la expresión:

$$V_{D,\text{sat}} = V_G - 2\psi_i + \frac{q\epsilon N_a}{C_{\text{ox}}^2} \left[1 - \left(1 + \frac{2V_G C_{\text{ox}}^2}{q\epsilon N_a} \right)^{1/2} \right]$$

- 8.10** Calcular el campo eléctrico en el óxido de puerta de un MOSFET de silicio de canal n, con espesor $t_{\text{ox}}=0.1 \mu\text{m}$ y longitud $L=10 \mu\text{m}$, cuando se aplica un voltaje de puerta $V_G=5.0 \text{V}$ y un voltaje de drenador $V_D=4.0 \text{V}$, en los siguientes casos: a) en un punto próximo a la fuente ($x=0$) y b) próximo al drenador ($x=L$). Comparar el resultado con el campo eléctrico necesario para la avalancha en una unión p-n.
- 8.11** Sea un transistor MOSFET de silicio de canal n, de longitud $L=1 \mu\text{m}$, anchura $z=20 \mu\text{m}$, capacidad del óxido por unidad de superficie $C'_{\text{ox}}=1.2 \times 10^{-3} \text{Fm}^{-2}$ y voltaje umbral $V_T=1.1 \text{V}$. Se pide calcular la corriente de saturación, $I_{D,\text{sat}}$, y la transconductancia, g_m , cuando el voltaje aplicado a la puerta es $V_G=6 \text{V}$. (Tómese para la movilidad de los electrones, $\mu_e=1000 \text{cm}^2\text{V}^{-1}\text{s}^{-1}$).
- 8.12** En un MOSFET de silicio de canal n, con $z=30 \mu\text{m}$, $L=1 \mu\text{m}$, $\mu_e=750 \text{cm}^2\text{V}^{-1}\text{s}^{-1}$, y $C'_{\text{ox}}=1.5 \times 10^{-3} \text{Fm}^{-2}$, calcular la corriente de saturación y la transconductancia para un voltaje aplicado $V_G=5 \text{V}$.

CAPITULO IX

APLICACIONES DE LOS TRANSISTORES COMO DISPOSITIVOS AMPLIFICADORES

Una de las aplicaciones más importantes de los transistores en electrónica analógica es la de amplificación de señales eléctricas de amplitud variable, tanto de voltaje como de corriente. Dependiendo de la función que se pretenda realizar, los circuitos amplificadores pueden ser de diversos tipos, tales como amplificadores de potencia, amplificadores sintonizados, etc. A su vez, los amplificadores constituyen la base de otros circuitos más complejos. Ejemplos típicos lo forman las diferentes familias de generadores de señal, las fuentes de alimentación, los amplificadores operacionales, etc. Dada la enorme amplitud de estos temas, en este capítulo nos centraremos solamente en los fundamentos de los circuitos de amplificación basados en los transistores bipolares y de efecto campo (FET). En capítulos posteriores se tratarán otros tipos de amplificadores algo más complejos con aplicaciones especiales.

9.1. CIRCUITOS AMPLIFICADORES

En la fig. 9.1a se presenta un esquema de los elementos básicos de un circuito amplificador simple para señales alternas. Este amplificador se puede considerar como un circuito de dos puertas. La puerta de entrada recibe la señal de voltaje o corriente que se pretende amplificar. Esta señal es procesada en el amplificador y es entregada a través de la puerta de salida convenientemente amplificada (o reducida en algunos casos). Cuando se trata de señales

pequeñas generalmente se supone que la señal de salida es semejante a la señal de entrada, es decir que no sufre ninguna distorsión, siendo la única diferencia la amplitud de la variación de la señal y la fase. Entre los parámetros que mejor definen el comportamiento de un circuito amplificador se encuentran la ganancia y las impedancias de entrada y salida. Veamos el significado de cada uno de estos factores.

Los términos *ganancia o factor de amplificación* se usan indistintamente para señalar la relación entre las amplitudes de las señales de entrada y salida. Así se define el factor de amplificación de voltaje, A_v , a través de la relación:

$$A_v = \frac{v_o}{v_i} \quad [9.1]$$

donde v_o y v_i representan, respectivamente, las amplitudes de los voltajes de la onda de salida y de entrada. Análogamente, para señales de corriente, el factor de amplificación en corriente, A_i , vendrá definido por:

$$A_i = \frac{i_o}{i_i} \quad [9.2]$$

siendo i_o e i_i las amplitudes de las corrientes de la onda de salida y entrada, respectivamente. Si en un circuito existe a la vez amplificación del voltaje y de la corriente de entrada, se puede definir también la ganancia en potencia, A_p , como la relación entre las potencias de la onda de salida y entrada, esto es $A_p = p_o / p_i$. De las relaciones anteriores, tendremos para la ganancia en potencia: $A_p = A_v \cdot A_i$. Es importante señalar que cuando existe desfase entre la onda de entrada y salida, el factor de amplificación queda determinado mediante dos términos, por lo que su valor suele ser expresado en forma de número complejo.

De los dos tipos de amplificadores señalados, en lo que sigue nos centraremos principalmente en los amplificadores de señales de voltaje, ya que éstos son los de mayor utilización práctica. Cuando un amplificador de voltaje se examina desde la puerta de entrada se puede considerar que está constituido por una resistencia (o una impedancia compleja, en el caso más general) entre los dos terminales de entrada que recogen la señal de voltaje, fig 9.1b. Esta resistencia se denomina impedancia de entrada del amplificador, R_i , y representa la carga para la señal de entrada aplicada al amplificador. El valor de R_i se puede calcular a través de la relación entre el voltaje de la señal y la corriente producida en el terminal de entrada, esto es: $R_i = v_i / i_i$. **Cuando se trata de amplificar señales de voltaje interesa que la impedancia de entrada del amplificador sea lo más alta posible con objeto de reducir al máximo la corriente i_i que el generador entrega al amplificador.**

Del mismo modo, considerando el amplificador como un circuito que entrega una señal de voltaje a la salida, se puede representar su comportamiento en lo que se refiere a los terminales de salida, de acuerdo con el teorema de Thévenin (véase apéndice A3), mediante

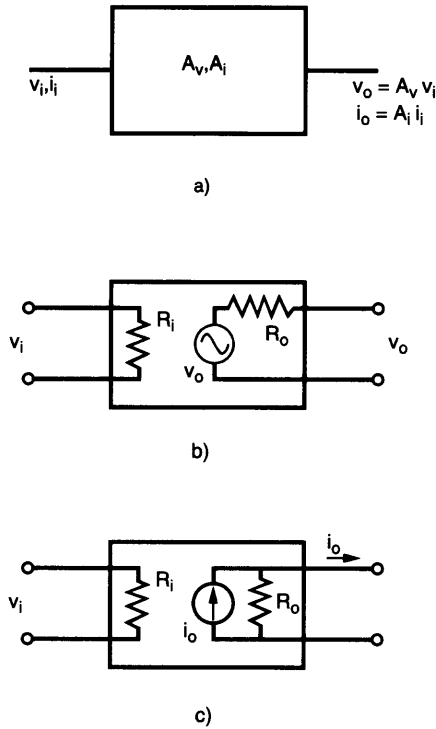


Fig. 9.1. a) Esquema de un circuito amplificador básico. b) Esquema del amplificador considerado como un circuito de dos puertas, con salida de señal de voltaje. c) Esquema del amplificador con salida de señal de corriente.

un generador de tensión en serie con una resistencia (fig.9.1b). Según este esquema, el generador de tensión ha de tener un valor igual a la tensión medida entre los terminales de salida en circuito abierto, esto es v_o . La resistencia en serie con el generador se denomina, en el caso más general, impedancia de salida, R_o , y su valor se obtiene dividiendo la tensión del generador por la corriente de cortocircuito en la salida (es decir, la obtenida al cortocircuitar los terminales de salida). **Desde el punto de vista del consumo de corriente en la resistencia que recibe la señal de salida, interesa que la impedancia de salida sea lo más baja posible ya que de esta forma el voltaje a la salida del generador, v_o , sufre una menor atenuación en la resistencia de salida.**

Alternativamente, en algunos casos es conveniente considerar el amplificador como un dispositivo que entrega en la salida una señal de corriente, i_o , proporcional al voltaje de entra-

da. El comportamiento del amplificador en lo que se refiere a los terminales de salida puede ser descrito entonces, siguiendo el teorema de Norton, mediante una fuente de corriente de intensidad i_o y una resistencia en paralelo de valor R_o . Esta resistencia está determinada por el cociente entre el voltaje medido en circuito abierto y la corriente en cortocircuito, i_o . En el caso más general, se trata de una impedancia en paralelo en lugar de una resistencia pura, y constituye la impedancia de salida del amplificador (fig. 9.1c). **Interesa ahora que esta resistencia paralelo sea lo más elevada posible con objeto de que el amplificador pueda entregar la señal de corriente sin ninguna atenuación en la propia resistencia interna, R_o .** Como veremos en el siguiente apartado, los circuitos amplificadores simples basados en un transistor bipolar conectado en la configuración de emisor común o de base común pueden considerarse, en lo que se refiere a la señal de salida, como una fuente de corriente.

9.2. EL TRANSISTOR BIPOLAR COMO AMPLIFICADOR

9.2.1. Circuito amplificador de base común.

La fig. 9.2a muestra el esquema de un circuito amplificador simple basado en un transistor bipolar, tipo pnp, conectado en la configuración de base común. Según se indica, el transistor está polarizado en la región activa mediante dos fuentes de alimentación, V_{EE} y V_{CC} ¹, que polarizan las uniones de emisor y de colector en directo y en inverso, respectivamente. A la entrada del circuito amplificador se ha incluido el generador de voltaje que es el que entrega la señal que se pretende amplificar, v_i . Esta señal se supone que es pequeña en comparación con V_{EE} y V_{CC} . Además, suponemos también que se trata en este caso de un generador ideal, y por tanto la resistencia serie del generador es nula. Asimismo, a la salida del amplificador se ha conectado una resistencia, R_L , que representa la resistencia de consumo, también denominada *resistencia de carga*, sobre la cual aparece la señal amplificada, v_o . Generalmente, esta resistencia se coloca en serie con la fuente de tensión V_{CC} , y por tanto sirve también para polarizar la unión de colector a una tensión V_{BC} , diferente a V_{CC} .

Para entender cómo funciona el circuito amplificador vamos a suponer primero que la señal del generador es nula. En este caso, la tensión V_{EE} aplicada en el circuito de entrada polariza en directo la unión de emisor a una tensión $V_{EB} = V_{EE}$ y produce una corriente I_E en el terminal de emisor. En un caso típico, la fuente de alimentación V_{EE} suele ser de unas décimas de voltio; con ello la corriente I_E puede ser de unos miliamperios (fig. 6.8). Por otra parte, en el circuito de salida la tensión V_{CC} da lugar a una cierta polarización inversa V_{BC} en la unión de colector, por lo que la corriente I_C en el terminal de colector vendrá dada aproximadamente

¹ **Nota:** Es práctica frecuente designar con doble subíndice en mayúscula la polarización en continua suministrada por las fuentes de alimentación, con objeto de distinguir estos voltajes de los que aparecen en los terminales del transistor.

por: $I_C \approx \alpha_{dc} I_E$, con $\alpha_{dc} \approx 1$ (véase sec. 6.2.3). El circuito en esta situación se encuentra en un estado de polarización estática, también denominada *quiescente*, con tensiones y corrientes constantes.

La presencia de una señal, v_i , en el circuito de entrada, no tiene por qué afectar el estado de polarización quiescente siempre que el valor de la señal sea muy pequeña en relación con la tensión de alimentación V_{EE} . La unión de emisor seguirá entonces polarizada en directo, con una tensión, v_{EB} , igual a la suma de v_i y V_{EB} . Debido a la presencia de la señal variable, v_i , en el terminal de emisor se produce también una corriente variable, i_e , superpuesta a la corriente I_E originada por la polarización en continua. La corriente i_e vendrá determinada por la ecuación $v_i = r i_e$, donde r es la resistencia dinámica de entrada del transistor en la confi-

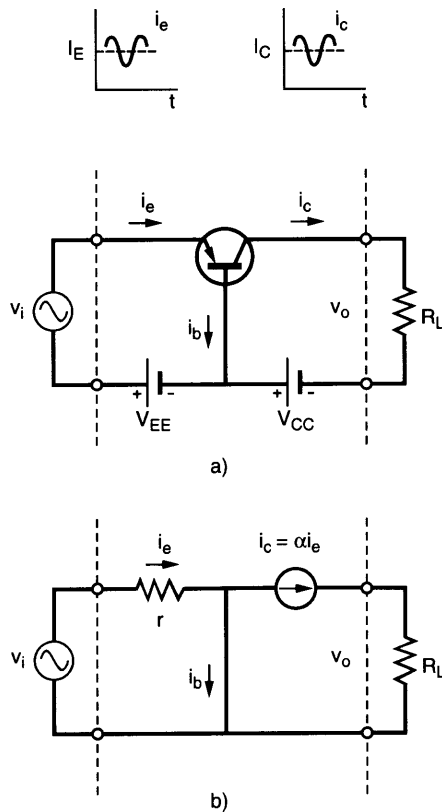


Fig. 9.2. a) Circuito amplificador simple formado por un transistor pnp en la configuración de base común, polarizado en la región activa. b) Circuito equivalente de pequeña señal (en la parte superior se representa la variación de las corrientes de emisor y de colector, respectivamente).

guración de base común (ec. 6.32). La señal de corriente i_c produce a su vez en la unión de colector una corriente variable, $i_c = \alpha i_e$, superpuesta a la corriente de continua del colector, I_C (α es el factor de ganancia en corriente para señales alternas dado por la ec. 6.8). Esta señal de corriente i_c circula por el circuito de salida del amplificador y da lugar a una caída de voltaje en la resistencia R_L cuyo valor, v_o , está dado por $v_o = i_c R_L$. Obviamente, esta caída de voltaje aparece también superpuesta a la caída de voltaje originada por la corriente continua del colector. Si R_L es suficientemente grande, la señal de salida, v_o , resulta amplificada en relación a la señal de entrada, v_i , y al mismo tiempo variará en fase con ella.

A partir de estas consideraciones es fácil obtener el factor de amplificación del circuito. Dado que se trata de un amplificador para señales de voltaje tendremos:

$$A_v = \frac{v_o}{v_i} = \frac{i_c R_L}{i_e r} = \alpha \frac{R_L}{r} \quad [9.3]$$

En esta ecuación, r representa la resistencia dinámica de la unión de emisor polarizada en directo, según se ha mencionado más arriba. En el rango de funcionamiento típico del transistor, esta resistencia tiene un valor muy pequeño, alrededor de unos 20 ohmios. Por otra parte, el valor de R_L se suele elegir bastante elevado, aunque menor que la resistencia de salida tipo paralelo asociada al circuito amplificador. Un valor típico para R_L puede ser alrededor de 5000 ohmios. Con esto, el factor de amplificación de voltaje resultaría del orden de 250.

Aunque el circuito de la figura 9.2 está diseñado para amplificar señales de voltaje, podemos suponer también que existe una amplificación de la corriente introducida en el circuito de entrada. De acuerdo con las definiciones del apartado anterior, el factor de amplificación en corriente vendrá dado por:

$$A_i = \frac{i_o}{i_i} = \frac{i_c}{i_e} = \alpha \quad [9.4]$$

ya que, según se desprende de la fig. 9.2a, $i_o = i_c$, e $i_i = i_e$. Dado que en la región activa $\alpha \approx 1$, podemos concluir que el circuito amplificador de la fig. 9.2 no presenta una amplificación apreciable para la corriente del circuito de entrada.

El cálculo de la resistencias de entrada y salida del circuito amplificador puede hacerse con ayuda del circuito equivalente representado en la fig. 9.2b. En esta figura, se ha sustituido el transistor por su circuito equivalente de señales alternas descrito en la fig. 6.17 (recuérdese que, por tratarse de señales alternas, las fuentes de alimentación en continua quedan cortocircuitadas en el circuito equivalente). El circuito contiene en la entrada una resistencia de valor r y en la salida un generador de corriente constante de valor $i_c = \alpha i_e$. Del esquema de la fig. 9.2b se desprende que la resistencia de entrada, R_i , coincide con la resistencia dinámica de la unión de emisor polarizada en directo, cuya magnitud hemos denominado r .

Del mismo modo, considerando el circuito de salida como un generador de señal de corriente i_c , la resistencia de salida, R_o , es una resistencia colocada en paralelo con el generador de corriente, tal como se indicó en la fig. 9.1c. Es fácil demostrar que el valor de esta resistencia se corresponde con el inverso de la pendiente de las curvas características de salida del transistor en el punto de operación, para la configuración de base común, fig. 6.9. Dado que en la región activa del transistor las curvas características son prácticamente horizontales, la resistencia R_o toma valores muy elevados y por esta razón no aparece representada en la fig. 9.2b.

Se concluye, pues, que el circuito amplificador de base común posee una resistencia de entrada relativamente baja. Esta característica es poco deseable para un amplificador ideal de señales de voltaje. Sin embargo, **un aspecto interesante de este circuito es que se comporta en lo que se refiere a la salida como una fuente ideal de señales de corriente, con una resistencia de salida bastante elevada, por lo que la señal de corriente de salida es prácticamente independiente del valor de la resistencia de carga.**

9.2.2. Circuito amplificador de emisor común

Veamos ahora cómo el transistor conectado en la configuración de emisor común puede ser utilizado como un circuito amplificador de gran versatilidad. En la fig. 9.3a se da el esquema de un circuito amplificador típico empleando un transistor tipo pnp polarizado en la región activa. La polarización en la región activa se consigue mediante dos baterías, V_{BB} y V_{CC} situadas en los circuitos de entrada y salida, respectivamente. Típicamente, V_{BB} ha de tener un valor de unas décimas de voltio (alrededor de 0.6-0.7 V) con objeto de que la corriente I_B en la unión de emisor sea del orden de los microamperios (véase fig. 6.10). En cambio V_{CC} se puede elegir con mayor flexibilidad, ya que su misión es mantener el colector suficientemente negativo respecto a la base. Recuérdese que, en la región activa, la corriente I_C es prácticamente independiente de la tensión V_{BC} . En el circuito de la fig. 9.3a se ha incluido, al igual que en el caso anterior, un generador de señales de voltaje, v_i , conectado a la entrada y una resistencia de carga, R_L , conectada a la salida para recibir la señal amplificada.

Para estudiar el comportamiento del circuito podemos suponer inicialmente que, al igual que en el caso anterior, la amplitud de la señal de entrada es cero. La tensión V_{BB} en el circuito de entrada produce entonces una tensión en directo $V_{EB} = V_{BB}$ en la unión de emisor, lo cual da lugar a una corriente I_B en el terminal de base. Al mismo tiempo, el potencial V_{CC} del circuito de salida polariza el transistor a una cierta tensión V_{EC} , de modo que el transistor opera en la región activa. En el circuito de salida tendremos una corriente I_C cuyo valor viene dado por $I_C \approx \beta_{dc} I_B$, siendo β_{dc} el factor de ganancia en corriente del transistor para la configuración de emisor común (definido a través de la ec. 6.9). Se dice que en estas circunstancias el transistor se encuentra también en un estado de polarización estática o quiescente.

Cuando la señal de entrada, v_i , es diferente de cero el estado de polarización del circuito no cambia siempre que el valor de la señal sea mucho menor que la tensión de alimentación del circuito de entrada, V_{BB} . De nuevo ocurre que la señal de entrada, v_i , produce una variación de corriente en el circuito de base de amplitud i_b , que se superpone a la corriente en continua que circula por la base, I_B . La corriente i_b está relacionada con el valor de la señal de entrada a través de la ecuación: $v_i = r' i_b$, siendo r' la resistencia dinámica de la unión de emisor para la configuración de emisor común (sec. 6.5.2). La señal i_b genera a su vez una variación de la corriente de colector, i_c , dada por $i_c = \beta i_b$ (β es el factor de ganancia en corriente alterna del transistor en la configuración de emisor común). La corriente i_c se superpone a la corriente en continua de colector, I_C , y da lugar a una caída de tensión en la resistencia de carga, $v_o = i_c R_L$, correspondiente a la señal de salida.

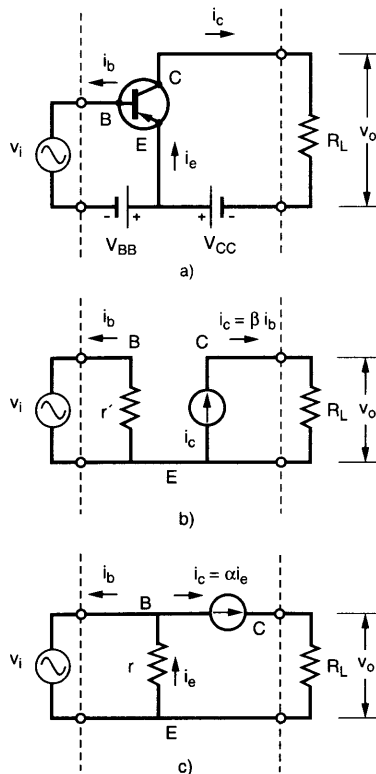


Fig. 9.3. a) Circuito amplificador formado por un transistor pnp en la configuración de emisor común, polarizado en la región activa. b) Circuito equivalente de pequeña señal, incluyendo la resistencia r' y el generador de corriente $i_c = \beta i_b$. c) Circuito equivalente alternativo, con la resistencia r y el generador de corriente $i_c = \alpha i_e$.

El factor de amplificación de voltaje para la configuración de emisor común será:

$$A_v = \frac{v_o}{v_i} = \frac{i_c R_L}{i_b r'} = \beta \frac{R_L}{r'} \quad [9.5]$$

Como se recordará (véase sec. 6.5.2), la resistencia r' correspondiente a la configuración de emisor común es mucho mayor que la de base común ($r' \approx \beta r$), con un valor alrededor de unos 2000 ohmios. Si por ejemplo $R_L = 800$ ohmios (valor típico) y $\beta = 100$ se tiene $A_v = 40$.

Así mismo, podemos considerar que el circuito produce también una amplificación en la corriente que produce la señal de voltaje en el circuito de entrada. El factor de amplificación en corriente del circuito será entonces:

$$A_i = \frac{i_o}{i_i} = \frac{i_c}{i_b} = \beta \quad [9.6]$$

Dado que el valor de β suele ser del orden de 100, podemos concluir que el factor de amplificación de corriente de este circuito es siempre muy elevado. Se aprecia, pues, que **el circuito de emisor común produce simultáneamente una amplificación tanto en la señal de voltaje como en la corriente de entrada, lo cual da lugar a una fuerte amplificación en potencia.**

El circuito equivalente de pequeña señal del amplificador viene mostrado en la fig. 9.3b. Este circuito se corresponde con el que ha sido desarrollado en la fig. 6.18a al estudiar el comportamiento del transistor en corriente alterna. Obsérvese que la resistencia de entrada del circuito amplificador, R_i , coincide con r' . La resistencia de entrada de este amplificador es, pues, mucho mayor que la del amplificador de base común, descrito anteriormente. Del mismo modo, el circuito de salida está formado por una fuente de corriente ideal, de valor $i_c = \beta i_b$ constante. Es fácil demostrar que este circuito coincide con el de la figura 6.18a, en el cual la fuente de corriente está dada por el valor $g_m v_{eb}$, siendo g_m la transconductancia del transistor (ec. 6.39). La resistencia de salida, R_o , asociada a la fuente de corriente constante es una resistencia paralelo de valor relativamente alto (no representada en la figura), ya que su magnitud coincide con el inverso de la pendiente de las curvas características del transistor en la configuración de emisor común (fig. 6.11). La pendiente de estas curvas es algo mayor que las de base común y por tanto la resistencia de salida es ahora menor que en el caso anterior.

Como veremos más adelante, en el análisis de algunos amplificadores más complejos es conveniente la utilización de otro circuito equivalente del amplificador, alternativo al anterior, que describe igualmente el comportamiento para señales alternas. Este circuito equivalente viene indicado en la fig. 9.3c, y ha sido descrito anteriormente en el estudio del transistor (fig. 6.18b). Según vimos, el circuito lleva asociado una resistencia r común a los circuitos de entrada y salida (r coincide aquí con la resistencia dinámica de la unión de emisor en la confi-

guración de base común). Al mismo tiempo, en el circuito de salida existe, al igual que el representado en la fig. 9.2b, una fuente de corriente constante en el circuito de salida de valor $i_c = \alpha i_e \approx i_e$.

Como resumen de la discusión anterior, podemos concluir que el amplificador de emisor común amplifica tanto la señal de voltaje de entrada como la corriente resultante en el circuito de entrada, resultando en una amplificación en potencia elevada. Posee asimismo una resistencia de entrada más alta que el de base común y su comportamiento en lo que se refiere al circuito de salida se puede asimilar al de una fuente de corriente constante con una resistencia de salida en paralelo de valor relativamente elevado. **Todas estas características son muy adecuadas para un amplificador y hacen que el transistor polarizado en la configuración de emisor común sea muy apropiado para la amplificación de señales alternas.**

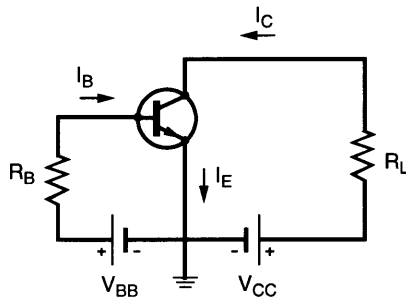


Fig.9.4. Circuito amplificador con un transistor npn en la configuración de emisor común, polarizado en la región activa. El circuito de polarización de base incorpora una resistencia R_B en serie con la batería V_{BB} .

9.3. DETERMINACION DEL PUNTO DE FUNCIONAMIENTO DEL TRANSISTOR EN EL CIRCUITO AMPLIFICADOR

Supongamos el circuito amplificador de la fig. 9.4 con un transistor npn en la configuración de emisor común², polarizado en la región activa mediante las fuentes de alimentación V_{BB} y V_{CC} . En lo que sigue nos referiremos siempre a la configuración de emisor común, ya que es la más utilizada en los circuitos amplificadores. El amplificador incluye, a diferencia del anterior (fig. 9.3a), una resistencia R_B en el terminal de base, con objeto de polarizar la unión de emisor a una tensión V_{BE} diferente a la tensión V_{BB} . Además, el terminal de emisor

² **Nota:** Por conveniencia, utilizaremos en lo que sigue transistores tipo npn, a diferencia de los circuitos anteriores en los que el transistor era del tipo pnp. Como veremos más adelante, esto facilita la asignación del sentido de las corrientes.

se halla unido a tierra con objeto de tener un punto de referencia de potenciales, común a los circuitos de entrada y salida. La conexión a tierra del emisor obliga a que tanto la señal de entrada como la de salida del amplificador queden también referidas al potencial de tierra.

Para estudiar el funcionamiento del amplificador es conveniente examinar primero su comportamiento en corriente continua y determinar al mismo tiempo el punto de funcionamiento del transistor. La determinación del punto de funcionamiento implica el cálculo de la tensión V_{BE} y la corriente I_B , correspondientes al circuito de entrada, así como la tensión V_{CE} y la corriente I_C en el circuito de salida.

9.3.1. Circuito de entrada: Determinación de la corriente de base

Como veremos más adelante, en las condiciones de polarización típicas del transistor en la región activa, la corriente de base I_B es la magnitud que puede tener una mayor variación en el circuito de entrada del amplificador. El cálculo de I_B así como de la tensión V_{BE} en el circuito de la fig. 9.4, se puede hacer mediante aplicación de la segunda ley de Kirchhoff al circuito de entrada. Tendremos en este caso:

$$I_B R_B + V_{BE} - V_{BB} = 0$$

Despejando I_B resulta:

$$I_B = \frac{V_{BB} - V_{BE}}{R_B} \tag{9.7}$$

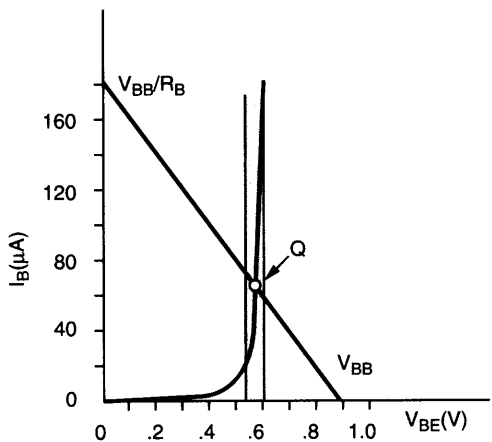


Fig. 9.5. Determinación del punto de funcionamiento del amplificador en el circuito de entrada. La zona sombreada muestra el rango de variación útil de la corriente I_B y el voltaje V_{BE} (0.6 V para el Si).

Si se representa I_B en función de V_{BE} se obtiene una recta de pendiente igual a $-1/R_B$ que corta a los ejes X e Y en los puntos V_{BB} y V_{BB}/R_B , respectivamente. En la fig. 9.5 se ha representado la recta dada por la ec. [9.7] para el caso particular de $V_{BB} = 0.9 \text{ V}$ y $R_B = 5 \times 10^3 \text{ ohmios}$.

Para determinar la corriente I_B es preciso conocer el valor de la variable V_{BE} en el punto de operación. Una segunda relación entre I_B y V_{BE} se puede obtener a partir de la curva intensidad-voltaje característica del transistor para el circuito de entrada. En la figura 9.5 se muestra la curva característica típica de entrada para un transistor bipolar tipo npn en la configuración de emisor común. En esta gráfica, la intersección de la recta dada por la ecuación anterior con la curva característica permite la determinación independiente de los valores de I_B y V_{BE} , y con ello del punto de funcionamiento, Q. En el ejemplo mostrado, las coordenadas del punto Q vienen dadas por $I_B \approx 60 \mu\text{A}$ y $V_{BE} \approx 0.6 \text{ V}$.

Hay que tener en cuenta, sin embargo, que en la región activa la unión de emisor está polarizada en directo, y por tanto el valor de V_{BE} suele ser pequeño (unas décimas de voltio). Además, debido al carácter cuasi-exponencial de la corriente de base (ver fig. 9.5), el valor de V_{BE} tiene una variación muy pequeña (incluso menor que una décima de voltio) dentro del rango de variación útil de la corriente de base. Por esta razón, se suele tomar en cálculos aproximados para V_{BE} un valor fijo cuando el transistor opera en la región activa. Así para los transistores de silicio V_{BE} se hace igual a 0.6 V y para los de germanio 0.2 V . Esta aproximación permite hacer una determinación rápida de la corriente I_B a través de la utilización directa de la ecuación [9.7].

9.3.2. Circuito de salida. Recta de carga estática

La determinación del punto de funcionamiento en el circuito de salida requiere el cálculo de la tensión V_{CE} y la corriente I_C . Aplicando de nuevo la segunda ley de Kirchhoff al circuito de salida de la fig. 9.4 tendremos ahora:

$$I_C R_L + V_{CE} - V_{CC} = 0$$

la cual se puede escribir también de la forma:

$$I_C = \frac{V_{CC} - V_{CE}}{R_L} \quad [9.8]$$

La ecuación [9.8] representa una recta en el diagrama $I_C - V_{CE}$, denominada *recta de carga estática*, y viene representada en la fig. 9.6 junto con las curvas características de un transistor típico. La pendiente de la recta de carga es igual a $-1/R_L$, mientras que los puntos de corte con los ejes X e Y vienen determinados por los valores V_{CC} y V_{CC}/R_L , respectivamente. En la figura se ha representado el caso particular de $V_{CC} = 30 \text{ V}$ y $R_L = 1000 \text{ ohmios}$.

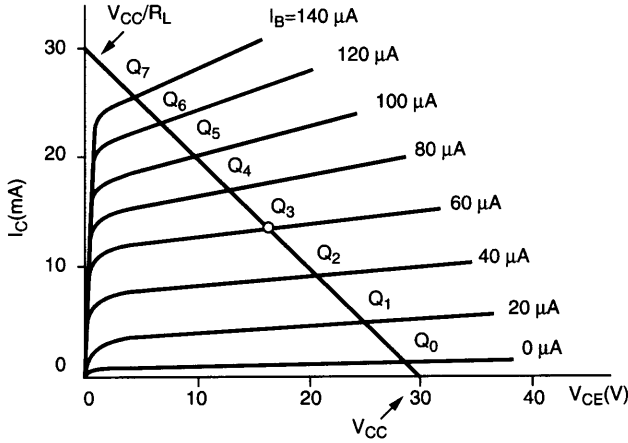


Fig.9.6. Recta de carga estática del circuito de la fig. 9.4, representada sobre las curvas características típicas de salida de un transistor npn.

Se observa que la recta de carga corta a las curvas características del transistor en una serie de puntos, Q_0 , Q_1 , Q_2 , etc. El punto de operación será el que se obtiene por intersección con la curva que corresponde a la corriente de base del transistor. En el ejemplo de la sección anterior, se obtuvo: $I_B \approx 60 \mu\text{A}$, por tanto el punto de operación corresponde al Q_3 de la fig. 9.6.

En electrónica analógica, este punto de trabajo, también denominado *punto quiescente*, se suele tomar cercano al centro de la línea de carga, de modo que las desviaciones alrededor de este punto, producidas por las señales alternas aplicadas en la base, sean lo más simétricas posible. En electrónica digital, por contra, el punto de trabajo se sitúa en los extremos de la línea de carga, esto es, en los puntos Q_0 y Q_7 , en los cuales la corriente del colector es prácticamente cero (punto Q_0) o bien toma un valor muy elevado (punto Q_7). De hecho, cuando se pretende que el punto de funcionamiento del transistor esté situado en una zona intermedia de la recta de carga, normalmente se opera en sentido inverso, es decir, primero se elige el valor adecuado de I_B en la curvas características, y después se calcula el valor correspondiente de R_B mediante aplicación directa de la ec. [9.7], suponiendo que el valor de V_{BE} es conocido, es decir, 0.6 V para los transistores de silicio.

9.3.3. Circuito amplificador con una fuente de alimentación única

El circuito de la fig. 9.4 requiere la utilización de dos fuentes de alimentación diferentes para mantener el estado de polarización del transistor en la región activa, lo cual plantea

algunos inconvenientes. Según hemos visto, cuando se introduce una resistencia en el circuito de base es posible polarizar la unión de emisor a una tensión diferente a la tensión de la fuente de alimentación, V_{BB} . Algo similar ocurre en el circuito de salida, en el cual la resistencia de carga permite que la tensión entre emisor y colector sea diferente a V_{CC} . De este modo, podemos decir que existe una cierta libertad de elección de los valores de V_{BB} y V_{CC} en el diseño del circuito amplificador. Es más, con objeto de simplificar el circuito, es posible elegir el mismo valor para ambas fuentes de alimentación. Con ello se tendría la ventaja adicional de que la polarización de los circuitos de entrada y salida del amplificador se podría tomar a partir de una fuente de alimentación única.

En la fig. 9.7 se presenta el esquema de un circuito amplificador de emisor común basado en estos conceptos. En este circuito, al igual que el de la fig. 9.4, el terminal de emisor se ha conectado a tierra para utilizar este punto como referencia de potenciales. Se ha tomado además una batería, V_{CC} , como fuente de alimentación única para ambos circuitos. Nótese que de este modo es posible polarizar las uniones de emisor y de colector en directo y en inverso, respectivamente, tal como se indica en la figura. Con objeto de simplificar el esquema del circuito, el polo positivo de la batería se ha situado en la parte superior del dibujo, separado del polo negativo (conectado a tierra). Así pues, se puede concluir que el circuito de polarización de la fig. 9.7 es para todos los efectos equivalente al de la figura 9.4, aunque por supuesto, el valor de la resistencia R_B es ahora diferente. Por razones que se harán aparentes más abajo, a la resistencia de carga se la denomina ahora resistencia de colector, R_C .

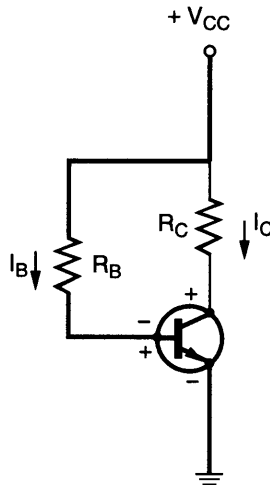


Fig. 9.7. Método simple de polarización de un transistor npn en la configuración de emisor común utilizando una sola fuente de alimentación, V_{CC} (la línea gruesa indica el circuito de entrada del amplificador).

Para determinar la corriente de base en el circuito del la fig. 9.7 se puede aplicar también la segunda ley de Kirchhoff al circuito de entrada, es decir el que contiene la resistencia R_B (delimitado en la fig. 9.7 con una línea de trazo grueso). Se tiene entonces para este circuito una ecuación análoga a la ec. [9.7], es decir:

$$I_B R_B + V_{BE} - V_{CC} = 0 \quad [9.9]$$

Siguiendo un procedimiento gráfico similar al señalado en la figura 9.5, es posible obtener los valores de I_B y de V_{BE} determinando el punto de intersección de la recta dada por la ec. [9.9] y la curva característica de entrada del transistor.

En el amplificador de la fig. 9.7, el circuito de salida tiene las mismas características que el amplificador de la figura 9.4 analizado anteriormente. De ahí que siga siendo válida la ecuación [9.8] para la determinación de la corriente de colector, aunque es preciso sustituir el valor de R_L en esta ecuación por el de la resistencia de colector, R_C . Así pues, se puede hacer la determinación del punto de funcionamiento del circuito de la fig. 9.7 siguiendo la misma rutina que en el circuito amplificador con dos fuentes de alimentación (fig. 9.4).

9.3.4. Acoplamiento de señales.

El funcionamiento del circuito amplificador en la configuración de emisor común requiere que la señal a amplificar, v_i , se introduzca en el terminal de base superpuesta a la tensión de polarización en continua de la unión de emisor, tal como se indicó en el esquema de la fig. 9.3a. Este requerimiento implica que en el circuito amplificador de la fig. 9.7 el generador de señal, v_i , debe ser conectado entre los terminales de base y emisor. Sin embargo, al ser la señal v_i pequeña, el acoplamiento directo de la señal hace que los terminales de base y emisor queden prácticamente en cortocircuito, llevando al transistor a la región de corte. Además el terminal del generador unido a la base quedaría a su vez conectado a la fuente de alimentación V_{CC} a través de la resistencia R_B , dando lugar a que una parte de la corriente que pasa a través de R_B circule también a través del generador de señal.

Para evitar estos problemas, normalmente se utiliza un condensador de acoplo, C_i , a la entrada del terminal de base. El terminal activo del generador de señal se conecta a este condensador con el otro extremo unido al emisor (potencial de tierra), tal como indica la fig. 9.8. Si la capacidad del condensador es elevada su impedancia capacitiva o reactancia, X_c , para señales alternas es relativamente baja y permite el paso de la señal alterna, ya que su valor viene dado por $X_c = 1/2\pi f C_i$ (f = frecuencia de la señal). En cambio, para la corriente continua el condensador ofrece una impedancia prácticamente infinita. Así pues, el condensador de acoplo permite superponer la señal v_i con la tensión continua de polarización en la base, manteniendo el transistor en la región activa, y al mismo tiempo impide que la corriente continua que pasa por R_B se derive hacia el generador de señal. En el circuito amplificador de la fig. 9.8 la señal amplificada de salida, v_o , se toma entre el terminal de colector y tierra

(nótese la diferencia con el circuito de la fig. 9.3a, en el que v_o se tomaba entre el colector y uno de los extremos de la batería). Esto tiene la ventaja de que la señal de salida queda de este modo referida a tierra, al igual que la señal de entrada.

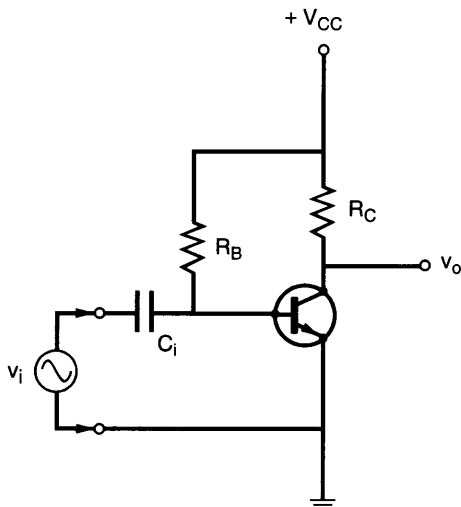


Fig. 9.8. Circuito amplificador para señales alternas utilizando un transistor npn en la configuración de emisor común, con un condensador de acoplo, C_i , para la señal de entrada.

Veamos ahora cómo afecta al estado de polarización estática la presencia de una señal de voltaje v_i aplicada a la entrada del circuito amplificador de emisor común con una fuente de alimentación única. Para visualizar el efecto de la señal de una forma gráfica, en la fig. 9.9 se ha representado de nuevo las curvas características de entrada y de salida del transistor, junto con la recta de carga. Supondremos que el punto de funcionamiento de polarización estática ha sido determinado previamente y corresponde al punto Q mostrado en las curvas. Tal como se ha indicado en la discusión del apartado anterior, la presencia del generador de voltaje, v_i , produce un voltaje alterno en la unión de emisor, v_{be} , que se superpone a la tensión continua aplicada en esta unión. Si por ejemplo la señal v_i es de 0.04 V pico a pico (p-p), la señal v_{be} será también de 0.04 V p-p y producirá una variación de la corriente I_B entre los puntos Q' y Q'', es decir: $i_b = 40 \mu\text{A}$ p-p, según se desprende de la fig 9.9a. Esta variación de la corriente de base se traduce a su vez en un desplazamiento del punto de funcionamiento a lo largo de la recta de carga alrededor del punto Q, esto es entre los puntos Q' y Q'' de la fig. 9.9b. El desplazamiento del punto de funcionamiento produce, asimismo, una variación en la corriente de colector, I_C , y también en la tensión entre emisor y colector, V_{CE} . La señal de corriente en

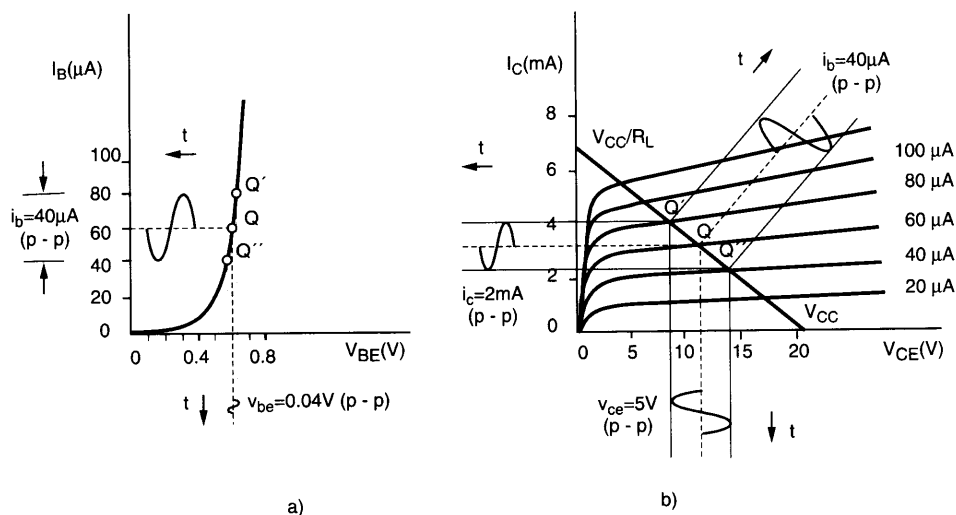


Fig. 9.9. Curvas características del circuito de entrada (a) y de salida (b) de un transistor npn en la configuración de emisor común, mostrando en cada caso las señales de voltaje y corriente en el transistor en función del tiempo.

el colector, i_c , será β veces mayor que la de i_b . En el ejemplo mostrado en la fig. 9.9b se observa que $i_c = 2 \text{ mA p-p}$ y $v_{ce} = 5 \text{ V p-p}$. Es importante señalar que las variaciones de i_b e i_c están en fase con la señal de entrada, v_i , mientras que la tensión v_{ce} oscila en oposición de fase (compárese la formas de la ondas i_b , i_c , v_{be} y v_{ce} mostradas en los diagramas de tiempo incluidos en la fig. 9.9). En el circuito de la fig. 9.4 la tensión de salida se tomaba entre los terminales de la resistencia R_L , con v_o dado por: $v_o = i_c R_L$, también oscilando en fase con la señal de entrada. Sin embargo, en el circuito de la fig. 9.8 la señal de salida coincide con la señal variable que existe entre los terminales de colector y emisor (tierra), esto es $v_o = v_{ce}$. Según se aprecia en la figura 9.9b, la tensión v_{ce} oscila en oposición de fase a la señal de entrada, v_i . Además, el cálculo del factor de amplificación de voltaje para este circuito arroja un valor que es también similar al obtenido anteriormente (ec. 9.5), excepto que ahora A_v toma un valor negativo, es decir: $A_v = -\beta(R_C/r')$. Podemos pues concluir que en el amplificador que estamos estudiando, con toma de la señal de salida entre el terminal de colector y tierra, **la señal de salida aparece amplificada respecto del valor de la señal de entrada y además desfasada en 180°** . Más adelante, al final de este capítulo, veremos cómo el condensador de acoplo C_i produce un desfase adicional en la señal v_i que entra en el transistor.

9.3.5. Efecto de la resistencia de carga: Recta de carga dinámica.

Cuando se desea entregar la señal de salida es necesario utilizar una resistencia de carga, R_L , externa al amplificador, conectada entre los terminales de colector y tierra. En este caso se hace preciso intercalar también en el terminal activo de salida del amplificador un condensador de acoplo, C_o , tal como se indica en el esquema de la fig. 9.10. La misión de este condensador es similar a la del condensador del circuito de entrada, esto es impedir que la corriente continua de la fuente de alimentación pase a través de R_L , produciendo un consumo innecesario. Así pues, con el condensador de bloqueo se aísla la carga del resto del amplificador y se evita con ello que la corriente continua pueda dañar la resistencia de carga. Obviamente, la capacidad del condensador debe ser suficientemente alta para conseguir que la señal de salida no se atenúe significativamente.

La presencia de una resistencia de carga, R_L , diferente a la resistencia de colector, R_C , así como del condensador de acoplo, C_o , obliga a reconsiderar el análisis efectuado en la determinación del punto de funcionamiento del circuito amplificador. Desde el punto de vista de la corriente continua, la presencia de la resistencia R_L no produce ningún efecto en el punto

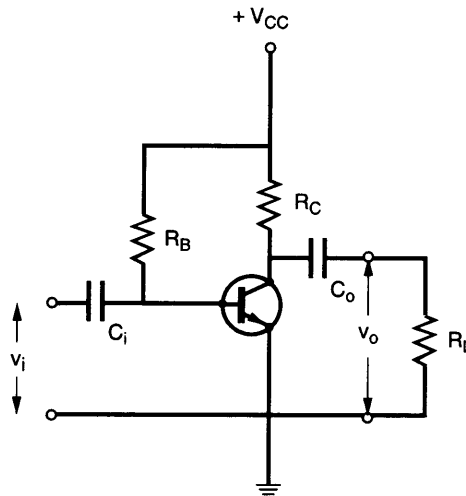


Fig. 9.10. Circuito amplificador para señales de alterna mostrando los condensadores de acoplo a la entrada y salida. Se incluye una resistencia de carga, R_L , a la salida del amplificador.

de funcionamiento determinado anteriormente (sec. 9.3.2), ya que esta resistencia está aislada para los efectos de la corriente continua del resto del circuito amplificador. Sin embargo, para las señales alternas, el condensador C_0 actúa como un cortocircuito, de modo que la resistencia de carga efectiva para la señal de corriente, i_c , viene dada por la asociación en paralelo de las resistencias R_C y R_L . Todo ello implica que en el análisis de las señales alternas la línea de carga deba ser modificada, ya que la nueva línea ha de tener una pendiente igual al inverso del valor obtenido al combinar en paralelo las resistencias R_C y R_L . Resulta así una recta de carga

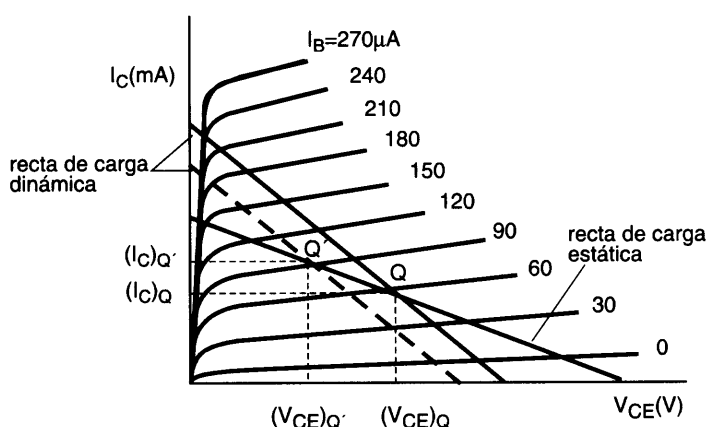


Fig. 9.11. Rectas de carga estática y dinámica para el circuito de la fig. 9.10.

en alterna con una pendiente mayor que la de continua. Como consecuencia de ello, la señal de salida v_o entre el terminal de colector y tierra es también diferente y toma ahora un valor proporcional a la resistencia resultante de la combinación de R_C y R_L en paralelo.

Es preciso hacer notar que la nueva línea de carga, también denominada *recta de carga dinámica*, debe pasar por el punto Q previamente determinado (en el ejemplo de la fig 9.11 se supone que al punto Q le corresponde una corriente $I_B = 60 \mu A$). La razón de que la recta de carga dinámica pase por el punto Q de funcionamiento, determinado bajo las condiciones de polarización estática, es que esta recta debe incluir entre sus puntos de funcionamiento el caso particular de señales con amplitud nula, lo cual coincide con el caso de polarización en corriente continua. Nótese que, en la recta de carga dinámica, el punto Q queda desplazado respecto al centro. El desplazamiento del punto Q hace que sea más conveniente situar el nuevo punto quiescente algo más arriba de la mitad de la recta de carga en continua, es decir con una corriente de base algo mayor (punto Q' por ejemplo, con $I_B = 90 \mu A$). La recta de carga dinámica debe trazarse paralela a la determinada anteriormente, pasando por el punto Q'

(línea a trazos en la fig. 9.11). De esta forma se obtiene una mayor variación posible para las señales de alterna.

9.4. ESTABILIDAD DEL PUNTO DE TRABAJO. CIRCUITO DE AUTOPOLARIZACION

El circuito de polarización de la corriente de base descrito en la sec. 9.3.1, aunque sencillo, no es demasiado práctico ya que mantiene el valor de I_B constante e independiente de las características del transistor (recuérdese que, según la ec. 9.9, $I_B \approx V_{CC}/R_B$). Esto puede resultar problemático en algunos casos, por ejemplo cuando hay que sustituir el transistor o también, caso frecuente, cuando hay variaciones de temperatura por calentamiento. En efecto, supongamos que un determinado transistor ha de ser sustituido por otro equivalente debido a que el primero se encuentra deteriorado. Generalmente, y de acuerdo con las hojas de especificaciones del transistor, ocurre que el valor de α_{dc} puede variar dentro de un pequeño margen, incluso para transistores del mismo tipo, esto es fabricados en las mismas condiciones. Sin embargo, pequeñas variaciones de α_{dc} pueden dar lugar a cambios bastante apreciables en el factor β_{dc} (recuérdese que $\beta_{dc} \approx \alpha_{dc}/(1-\alpha_{dc})$, con $\alpha_{dc} \approx 1$). Por este motivo, es común encontrar variaciones del valor de β_{dc} hasta en un 100 %. En estos casos, y dado que $I_C \approx \beta_{dc}I_B$, se tendrá que el espaciado de las curvas características de salida trazadas para ΔI_B constante variará proporcionalmente al cambio producido en β_{dc} . En el ejemplo de la fig. 9.12 se ha supuesto que β_{dc} aumenta respecto al valor original en un 20 % (curvas a trazos), por lo que el punto de trabajo sufre un desplazamiento desde Q hasta Q', con un valor de I_C más elevado.

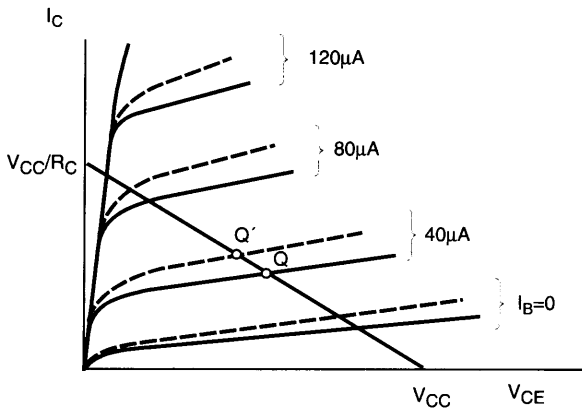


Fig. 9.12. Curvas características de salida para dos transistores del mismo tipo, pero con diferente valor de β_{dc} . Las líneas de trazos corresponden al transistor con mayor valor de β_{dc} .

En muchos casos el nuevo punto de funcionamiento puede resultar poco satisfactorio, e incluso puede ocurrir que el transistor entre en la región de saturación, con lo cual se pierde la linealidad de las curvas características. De esta discusión se desprende que la constancia de I_B no se traduce necesariamente en una mayor estabilidad del punto de trabajo al cambiar el valor de β_{dc} , como en principio se debería esperar. Por el contrario, es preferible que I_B varíe de modo que tanto I_C como V_{CE} (coordenadas del punto de trabajo) se mantengan prácticamente constantes dentro del rango de variación de β_{dc} .

Aparte de las variaciones producidas por eventuales cambios del valor de β_{dc} , el punto de trabajo también puede desplazarse como consecuencia de inestabilidades térmicas, debidas a cambios de temperatura. En efecto, las corrientes de base y de colector están relacionadas a través de la ecuación (véase apartado 6.2.3):

$$I_C = \beta_{dc} I_B + (\beta_{dc} + 1) I_{CBO} \quad [9.10]$$

donde I_{CBO} expresa la corriente en la unión de colector polarizada en inverso (corriente inversa de saturación). Esta corriente, que depende de la concentración de electrones y huecos minoritarios en las regiones p y n, respectivamente, es muy sensible a la temperatura, siendo normal que se duplique cada vez que la temperatura aumente en unos 10°C . Así pues, un aumento de la temperatura ambiente, o incluso el calentamiento producido por el paso de la corriente del colector, puede inducir un aumento de I_{CBO} . Este aumento a su vez provoca un incremento de I_C , lo que da lugar a un mayor incremento de la temperatura y hace que I_{CBO} sea progresivamente más elevada. Este proceso de realimentación puede llevar finalmente a la ruptura térmica del dispositivo. En otros casos, aunque no se produzca la ruptura térmica del dispositivo, el punto de trabajo se desplaza hacia arriba a lo largo de la recta de carga a medida que aumenta la temperatura, llegando a alcanzar la región de saturación siempre que I_B se mantenga constante durante el proceso.

9.4.1. Circuito de polarización universal o autopolarización

Uno de los métodos más utilizados para mantener el punto de trabajo estable es el denominado *circuito de polarización universal o autopolarización*, mostrado para la configuración de emisor común en la fig. 9.13a. En este circuito de polarización, el voltaje aplicado a la base se mantiene en un valor fijo, V_B , a través de un divisor de tensión formado por el conjunto de resistencias R_1 y R_2 . El circuito de estabilización incluye además una resistencia, R_E , conectada entre el terminal de emisor y tierra. La razón por la que se obtiene una mejora en la estabilidad es debido a que si por cualquier circunstancia I_C tiende a crecer, la corriente I_E a través de R_E también aumenta y por tanto también lo hace la caída de tensión a través de R_E . Como consecuencia de ello, la tensión en el emisor, V_E se vuelve más elevada, con lo que la tensión de la unión de emisor, $V_{BE} = V_B - V_E$, disminuye. Esta disminución de V_{BE} da lugar a que I_B disminuya también, compensando la tendencia inicial de I_C a aumentar (recuérdese

que $I_C \approx \beta_{dc} I_B$). De este modo, I_C se mantiene relativamente estable, incluso cuando existen cambios apreciables en la temperatura.

La clave de la estabilidad del punto de trabajo reside en que la tensión en el terminal de base, V_B , ha de mantenerse en un valor fijo, esto es, independiente de las variaciones en la corriente I_B . Esto se consigue siempre que la corriente a través del divisor de tensión formado por las resistencias R_1 y R_2 sea mucho mayor que la corriente de base, I_B . En estas circunstancias, las variaciones de I_B tendrán un efecto muy pequeño en la tensión V_B obtenida en un punto intermedio del divisor de tensión. Además, la resistencia R_E normalmente se elige mucho mayor que la resistencia de la unión de emisor en el punto de operación, r' . Con ello se consigue que las variaciones en r' (y por tanto en la tensión V_{BE}) debidas a posibles cambios de temperatura produzcan un menor efecto en la tensión V_B . En cualquier caso, el valor de R_E ha de mantenerse dentro de ciertos límites ya que, como veremos en el siguiente capítulo, la presencia de R_E en el circuito de amplificación reduce sensiblemente la ganancia del amplificador como consecuencia de un efecto de realimentación negativa (véase apartado 11.3).

Para analizar el circuito de la fig. 9.13a es conveniente sustituir el circuito que está a la izquierda de los terminales B y T (base y tierra, respectivamente) por su equivalente obtenido según el teorema de Thévenin (véase apéndice A3). De acuerdo con este teorema el circuito equivalente está formado por una fuente de alimentación, que denominaremos V_{BB} , en serie con una resistencia, R_B (fig. 9.13b). El valor de V_{BB} es igual al voltaje en circuito abierto entre los terminales B y T. Esto quiere decir que en el cálculo de V_{BB} el punto B ha de ser desconectado de la base. Análogamente, R_B viene dado por la resistencia que existe entre ambos terminales cuando se cortocircuitan las fuentes de alimentación (es decir, con V_{CC} conectado a tierra). De acuerdo con estos postulados, los valores de V_{BB} y R_B en el circuito de la fig. 9.13b

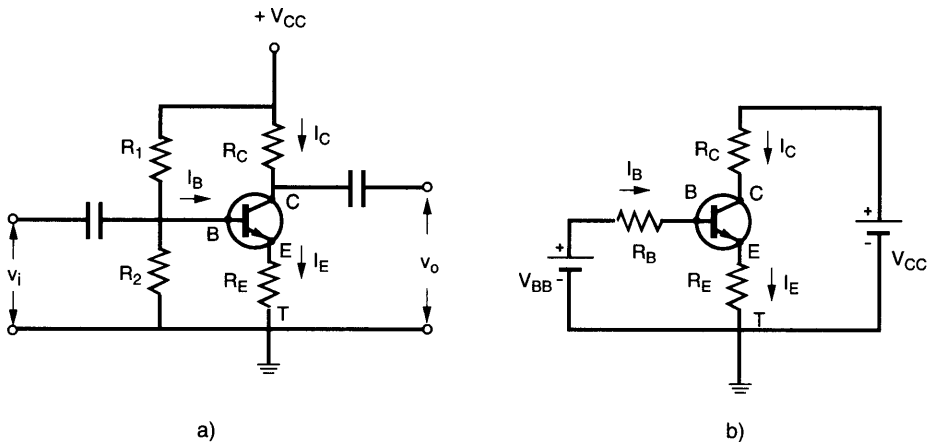


Fig. 9.13. a) Circuito de polarización universal. b) Circuito equivalente de Thévenin de polarización de base.

vienen dados por:

$$V_{BB} = \frac{R_2}{R_1 + R_2} V_{CC} \quad [9.11]$$

$$R_B = \frac{R_1 R_2}{R_1 + R_2} \quad [9.12]$$

La obtención del punto de trabajo en el circuito con autopolarización es algo más complejo que en los casos anteriores. Comencemos por aplicar la segunda ley de Kirchhoff a los circuitos de entrada y salida del amplificador. Tendremos:

$$V_{BB} = R_B I_B + V_{BE} + R_E (I_B + I_C) \quad [9.13]$$

$$V_{CC} = R_C I_C + V_{CE} + R_E (I_B + I_C) \approx V_{CE} + (R_E + R_C) I_C \quad [9.14]$$

En la última ecuación se ha hecho uso de la aproximación $I_B \ll I_C$ en la región activa. La ec. [9.14] representa la recta de carga en el circuito de salida del amplificador. Esta recta, junto con las curvas de salida del transistor, viene mostrada en la fig. 9.14. Según se indica, la

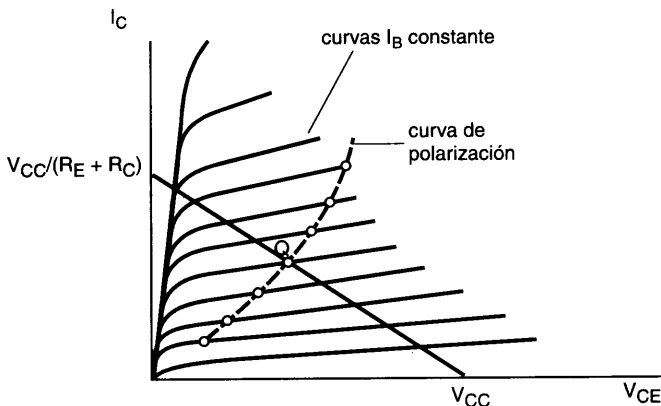


Fig. 9.14. Determinación del punto de funcionamiento mediante la intersección de la recta de carga con la curva de polarización.

recta corta a los ejes vertical y horizontal en los puntos $I_C = V_{CC}/(R_C + R_E)$ y $V_{CE} = V_{CC}$, respectivamente, y tiene una pendiente igual a $-1 / (R_C + R_E)$. Para determinar la corriente I_B correspondiente al punto de trabajo se puede despejar el valor de I_C de la ec. [9.14] y sustituirlo en la ec. [9.13], con lo que se obtiene una relación entre I_B y V_{CE} (el valor de V_{BE} se puede fijar de acuerdo con los criterios señalados en la sec. 9.3.1). Esta relación permite asignar un valor de V_{CE} para cada uno de los valores de I_B correspondientes a las curvas características del transistor. Si sobre cada una de ellas marcamos el punto cuya abscisa coincide con el valor de V_{CE} resultante, se obtiene una familia de puntos que forma la denominada *curva de polarización*. La intersección de la línea de polarización con la recta de carga determina finalmente el punto Q de trabajo, según se indica en la fig. 9.14.

Cuando se conoce la β_{dc} del transistor (a veces es posible extraer este valor directamente de las curvas características de salida del transistor) se puede obtener el punto de trabajo a partir de las ecs. [9.13] y [9.14] por un procedimiento más sencillo, ya que entonces es posible determinar simultáneamente V_{CE} , I_B e $I_C \approx \beta_{dc} I_B$.

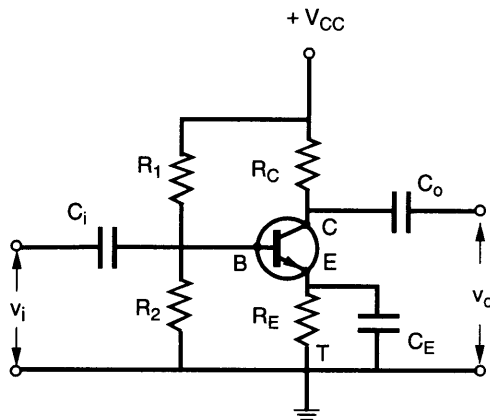


Fig. 9.15. Circuito amplificador completo, basado en un transistor npn en la configuración de emisor común. En el circuito se muestra el condensador de emisor, C_E , cuyo objeto es cortocircuitar el emisor a tierra para las señales alternas.

A menudo el circuito de autopolarización de emisor incluye también un condensador de gran capacidad, C_E , colocado en paralelo con la resistencia de emisor, tal como se indica en la fig. 9.15. Este condensador tiene por objeto cortocircuitar la resistencia R_E para las señales de alterna. Con ello se evita que exista el efecto de realimentación negativa de la señal de salida hacia el circuito de entrada y se consigue así que el factor de amplificación sea más elevado.

9.4.2. Factores de estabilidad en el punto de trabajo (*)

En apartados anteriores hemos visto que cuando varía la temperatura ocurre un cambio tanto en la corriente I_C como en el voltaje V_{CE} correspondiente al punto de trabajo. Así se encuentra por ejemplo que, en los amplificadores sin circuito de autopolarización, es decir con corriente de base constante, un aumento de la temperatura de 25 a 65 °C da lugar a un aumento del 100% en la corriente del colector. Según vimos, esta variación se debe fundamentalmente a que I_C es una función de β_{dc} e I_{CBO} y estos parámetros varían a su vez con la temperatura. Con objeto de analizar la influencia de la temperatura y del factor de ganancia, β_{dc} , en la corriente de colector cuando se utiliza el circuito de autopolarización es conveniente desarrollar una ecuación general que ligue estos parámetros. En efecto, podemos escribir la ec. [9.13] como:

$$V_{BB} = I_B R_B + V_{BE} + I_E R_E \quad [9.15]$$

El efecto de la temperatura sobre I_{CBO} y por tanto su influencia en la corriente de colector se puede obtener a partir de las relaciones obtenidas en el capítulo 6. Así, escribiendo de nuevo la ec. [6.16] tendremos:

$$I_C = \alpha_{dc} I_E + I_{CBO} \quad [9.16]$$

con

$$\alpha_{dc} = \beta_{dc} / (\beta_{dc} + 1) \quad [9.17]$$

Sustituyendo en la ec. [9.15] los valores de I_B e I_E obtenidos a partir de las ecs. [9.10] y [9.16] se obtiene después de operar:

$$I_C = \frac{\beta_{dc} (V_{BB} - V_{BE}) + I_{CBO} (\beta_{dc} + 1) (R_B + R_E)}{R_B + (\beta_{dc} + 1) R_E} \quad [9.18]$$

Es conveniente introducir los *factores de estabilidad*, S , del transistor para los cambios de I_C debidos a las variaciones de I_{CBO} , V_{BE} y β_{dc} en el transistor, a través de las ecuaciones:

$$S_{I_{CBO}} = \left. \frac{\Delta I_C}{\Delta I_{CBO}} \right|_{V_{BE}, \beta} = \frac{(\beta_{dc} + 1) (R_B + R_E)}{R_B + (\beta_{dc} + 1) R_E} \quad [9.19]$$

$$S_{V_{BE}} = \left. \frac{\Delta I_C}{\Delta V_{BE}} \right|_{I_{CBO}, \beta} = \frac{-\beta_{dc}}{R_B + (\beta_{dc} + 1) R_E} \quad [9.20]$$

$$S_{\beta} = \left. \frac{\Delta I_C}{\Delta \beta} \right|_{I_{CBO}, V_{BE}} = \frac{I_C - I_{CBO}}{\beta_{dc}} \left[\frac{R_B + R_E}{R_B + (\beta_{dc} + 1) R_E} \right] \quad [9.21]$$

los cuales indican la estabilidad de la corriente de colector, en el sentido de que cuanto menor sea S mayor es la estabilidad de la corriente.

Los cambios en la corriente de colector con relación al punto de trabajo debidos a las variaciones de I_{CBO} y V_{BE} originadas por la temperatura así como los debidos a las variaciones del factor β_{dc} se podrán expresar en función de los factores S como:

$$\Delta I_C = S_{I_{CBO}} \Delta I_{CBO} + S_{V_{BE}} \Delta V_{BE} + S_{\beta} \Delta \beta \quad [9.22]$$

Para utilizar las fórmulas anteriores se considera que para los transistores de silicio y de germanio V_{BE} decrece alrededor de 2.5 mV por cada grado centígrado. Es fácil demostrar además a partir de un análisis detallado de las expresiones [9.19] a [9.22] que los factores de estabilidad son mucho menores cuando existe circuito de autopolarización (es decir con R_E diferente de cero) que cuando no existe dicho circuito ($R_E = 0$).

9.5. AMPLIFICADORES CON TRANSISTORES DE EFECTO CAMPO DE UNION

Para los transistores de efecto campo se pueden diseñar circuitos amplificadores con características muy similares a las señaladas para los transistores bipolares. En este sentido, es conveniente recordar que la región de saturación de un transistor JFET juega un papel similar al de la región activa de los transistores bipolares. La configuración que se utiliza más a menudo es la de fuente común, esto es con el terminal de fuente común a los circuitos de entrada y salida del amplificador. Como veremos más adelante, los amplificadores basados en transistores de efecto campo, cuando están polarizados en la región de saturación, pueden ofrecer una relación casi lineal entre las señales de salida y entrada.

En la fig. 9.16 se presenta un circuito amplificador típico para un transistor de unión (JFET) de canal n. El circuito está alimentado mediante una fuente de continua V_{DD} que polariza conjuntamente los terminales de puerta y drenador (este último a través de la resistencia R_D). Como se recordará (véase apartado 8.2), en los JFET de canal n es preciso mantener el drenador a una tensión elevada respecto de la fuente con objeto de que el transistor opere en la región de saturación. Al mismo tiempo, el electrodo de puerta ha de estar polarizado negativamente respecto del electrodo de fuente. Esto se puede conseguir eligiendo adecuadamente las resistencias R_1 y R_2 del divisor de tensión que polariza la puerta. El circuito contiene

además una resistencia R_S entre el terminal de fuente o surtidor y tierra, la cual junto con el divisor de tensión sirve para estabilizar el punto de funcionamiento del transistor, según veremos más abajo. Esta última resistencia lleva asociada en paralelo un condensador, C_S , de gran capacidad cuyo papel es el de cortocircuitar la resistencia R_S para las componentes alternas de la señal (véase sec. 9.4.1). El circuito amplificador se completa con los condensadores de acoplo, C_i y C_o , para las señales de entrada y salida, respectivamente.

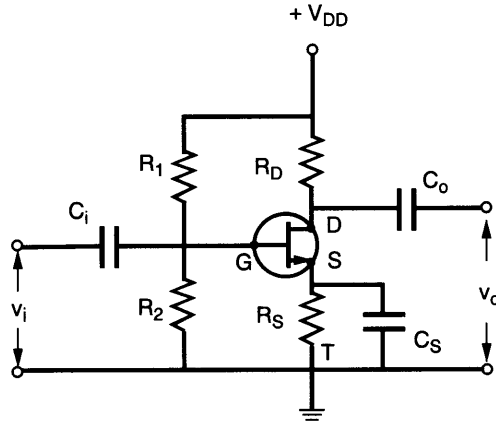


Fig. 9.16. Circuito amplificador basado en un transistor de efecto campo de unión de canal n .

La estabilización del punto de funcionamiento se consigue mediante el divisor de tensión y la resistencia R_S a través de un mecanismo similar al de los transistores bipolares. Esto es posible gracias al divisor de tensión que permite mantener el potencial de puerta, V_G , a una tensión fija, independientemente de las variaciones en la corriente de colector. Efectivamente, supongamos que en el circuito de la fig. 9.16 existe un aumento de la corriente de drenador, I_D (debida por ejemplo a un aumento de la temperatura). Este aumento de I_D se traduciría en una mayor corriente a través de la resistencia R_S y daría lugar a una mayor caída de potencial en esta resistencia. Con ello el potencial en el surtidor, V_S , se haría más elevado y la diferencia de potencial entre puerta y surtidor, $V_{GS} = V_G - V_S$ sería más negativa (recuérdese que V_G es negativo respecto al surtidor). La disminución de V_{GS} daría lugar a una disminución de la corriente de drenador, y con ello se compensaría la subida inicial de I_D .

Para determinar el punto de funcionamiento del transistor es preciso realizar un análisis siguiendo la segunda ley de Kirchhoff para los circuitos de entrada y salida. Así, para analizar el circuito de entrada es preciso sustituir primero el circuito que está a la izquierda de la puerta (G) y el terminal de tierra (T) por el circuito equivalente de Thévenin, tal como se hizo en el circuito de autopolarización del transistor bipolar (fig. 9.13). El circuito equivalen-

te está ahora formado por una fuente de alimentación de tensión, V_{GG} , cuyo valor viene dado por: $V_{GG} = [R_2/(R_1+R_2)]V_{DD}$, y una resistencia en serie, R_G (resistencia de Thévenin) de valor: $R_G = R_1R_2/(R_1 + R_2)$. Para el circuito de entrada tendremos entonces:

$$V_{GS} + I_D R_S - \frac{V_{DD} R_2}{R_1 + R_2} = 0 \quad [9.23]$$

En la ecuación anterior no se ha tenido en cuenta la caída de tensión en la resistencia equivalente R_G debido a que la corriente que circula por el terminal de puerta se puede considerar despreciable para efectos prácticos. Si en la ecuación anterior despejamos el valor de I_D en función de V_{GS} tenemos la denominada *recta de polarización*:

$$I_D = \frac{V_{DD} R_2}{(R_1 + R_2) R_S} - \frac{V_{GS}}{R_S} \quad [9.24]$$

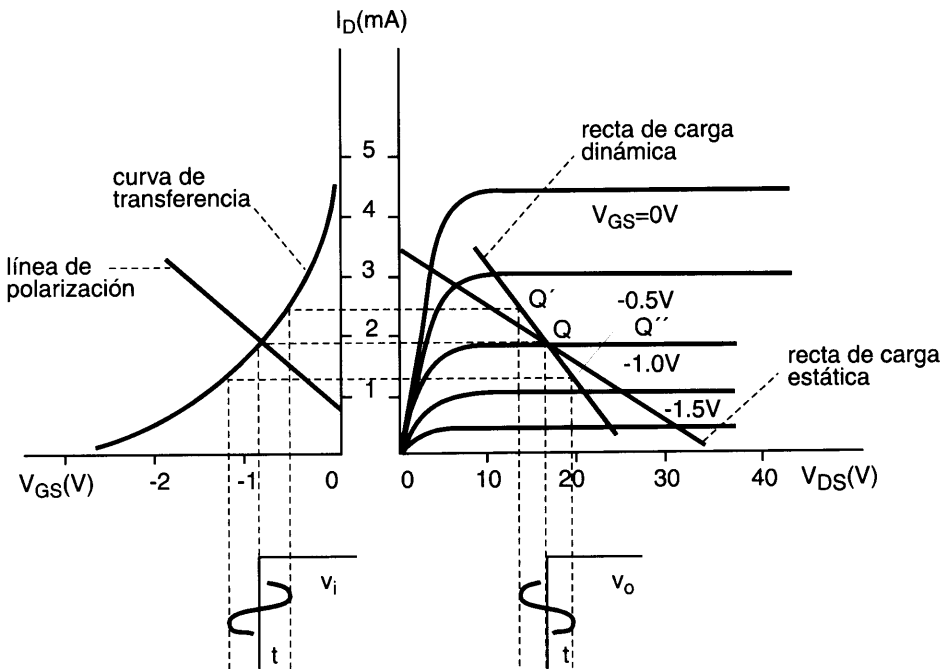


Fig. 9.17. Curvas características del circuito de entrada y salida de un transistor de efecto campo, empleadas en la determinación del punto de funcionamiento del amplificador de la figura anterior.

Análogamente, la aplicación de la segunda ley de Kirchhoff al circuito de salida permite escribir la siguiente ecuación para la recta de carga:

$$I_D = \frac{V_{DD}}{R_S + R_D} - \frac{V_{DS}}{R_S + R_D} \quad [9.25]$$

siendo V_{DS} la caída de tensión entre drenador y fuente.

Las ecs. [9.24] y [9.25] junto con las curvas características del transistor determinan el punto de funcionamiento. Así, en la parte izquierda de la figura 9.17 se ha representado *la curva de transferencia* del transistor (véase apartado 8.2) junto con la recta de polarización, dada por la ec. [9.24]. A la derecha de la figura se ha representado sobre las curvas características de salida del transistor la recta de *carga estática* (correspondiente a la ec. 9.25). Esta recta de carga corta a los ejes X e Y en los puntos V_{DD} y $V_{DD}/(R_S + R_D)$, respectivamente, y su pendiente viene determinada por el inverso de $R_S + R_D$. El punto de funcionamiento viene dado por la intersección de la recta de polarización con la curva de transferencia en la gráfica de la izquierda. En el ejemplo de la fig. 9.17, este punto tiene de coordenadas: $I_D \approx 1.8 \text{ mA}$ y $V_{GS} \approx -1 \text{ V}$. Llevando estos valores a la recta de carga del circuito de salida se obtiene el punto Q (punto quiescente) de funcionamiento de las curvas de salida.

La fig. 9.17 incluye también la recta de *carga dinámica* (véase apartado 9.3.5), la cual difiere de la recta de carga estática debido a la presencia del condensador C_s (y en su caso, del condensador C_o) conectado entre los extremos de la resistencia de surtidor, R_S . Este condensador sirve para cortocircuitar la resistencia R_S a las frecuencias normales de trabajo del amplificador y evitar con ello la caída de la ganancia asociada a la presencia de R_S . Lógicamente, la pendiente de la recta de carga dinámica será mayor que la estática, ya que viene dada por el inverso de la resistencia R_D solamente (o en su caso por la resultante de la asociación en paralelo de las resistencias R_D y R_L , si es que se añade una resistencia de carga al circuito).

En la fig. 9.17 se muestra el efecto obtenido al aplicar una onda sinusoidal en la entrada, $v_i = v_{gs}$, superpuesta a la tensión de polarización de puerta, V_{GS} . La señal $v_i = 0.6 \text{ V p-p}$ aparece representada en la parte inferior de la figura. Esta señal produce una variación en la corriente de drenador, i_d , del orden de 1 mA p-p , según se observa en la gráfica de la izquierda en la fig. 9.17. La señal de corriente i_d a su vez se traduce en un desplazamiento del punto de funcionamiento en la recta dinámica de carga, entre los puntos Q' y Q'', tal como se muestra en la gráfica de la derecha. Este desplazamiento del punto de funcionamiento origina finalmente una variación en la tensión V_{DS} , esto es $v_{ds} \approx 7.0 \text{ V}$, la cual coincide con la señal de salida, v_o . La forma de la señal de salida se ha representado también en la parte inferior de la figura. Según se aprecia, **esta onda aparece amplificada y al mismo tiempo tiene una variación en oposición de fase con la onda de entrada.**

En la fig. 9.17 puede observarse también que la amplitud de la onda de salida está determinada fundamentalmente por la pendiente de la recta dinámica de carga, cuya pendiente

es mayor que en el caso estático, según hemos visto. Además se cumple: $v_o = v_{ds} = -i_d R_D$. Por tanto, el factor de amplificación para señales pequeñas de voltaje se puede calcular a través de la ecuación:

$$A_v = \frac{v_o}{v_i} = \frac{v_{ds}}{v_{gs}} = - \frac{i_d R_D}{v_{gs}} = - g_m R_D \quad [9.26]$$

donde g_m es la transconductancia del transistor de efecto campo definida por la ec. [8.13]. El signo menos del resultado da cuenta del desfase entre las señales de entrada y salida.

De la discusión precedente se sigue que el circuito amplificador de fuente común basado en el JFET tiene unas características muy similares al circuito correspondiente del transistor bipolar. De hecho, el mecanismo de amplificación, la forma de las curvas características de salida y el factor de amplificación, son prácticamente idénticos a los descritos en el apartado anterior para el transistor bipolar. **Una característica esencial del circuito basado en el JFET, y que es común a los transistores de efecto campo es que la resistencia de entrada es muy elevada.** Según se comentó anteriormente (cap. 8), esto es debido a que la unión p-n que forma la puerta del transistor está polarizada en inversa. La alta resistencia de entrada hace que este amplificador sea especialmente aconsejable como primera etapa de amplificación para elevar el nivel de voltaje de una señal procedente de un generador con resistencia serie elevada. Sin embargo, en este tipo de amplificadores el factor de amplificación de señales de voltaje no es muy elevado (alrededor de 50) por lo que suelen ir seguidos de un amplificador basado en transistores bipolares que confiere al circuito una ganancia global más elevada.

9.6. AMPLIFICADORES CON TRANSISTORES TIPO MOSFET

Los transistores MOS de efecto campo también pueden ser utilizados en circuitos de amplificación. De hecho, las características intensidad-voltaje de los MOSFET son muy similares a las de los transistores de efecto campo de unión, según se ha señalado en el capítulo anterior. Por este motivo, los circuitos amplificadores basados en los MOSFET tienen también una estructura similar a los que se han presentado en las secciones precedentes, es decir el transistor se polariza, usualmente, en la configuración de fuente (surtidor) común. En estas circunstancias, el transistor debe operar en la región de saturación con objeto de obtener una relación aproximadamente lineal entre la onda de entrada y la de salida. Quizás sea conveniente recordar que los MOSFET de canal n operan en la región de saturación (equivalente a la región activa de un transistor bipolar) cuando el drenador está polarizado respecto de la fuente a un potencial, V_{DS} , positivo y de valor mayor que $V_{D,sat}$, mientras que la puerta se mantiene también a una tensión V_{GS} positiva, superior a un valor umbral (véase sec. 8.6.2).

En la fig. 9.18a se da un esquema de un circuito amplificador simple basado en un MOSFET de canal n, conectado según la configuración de fuente común. En este circuito, el

drenador está polarizado positivamente mediante la fuente de alimentación V_{DD} , conectada al drenador a través de la resistencia R_D . Asimismo, la puerta del transistor está polarizada positivamente mediante la resistencia R_F que une el terminal de puerta con el del drenador. Se ha incluido además los condensadores de acople de señales alternas en los terminales de entrada y salida del amplificador.

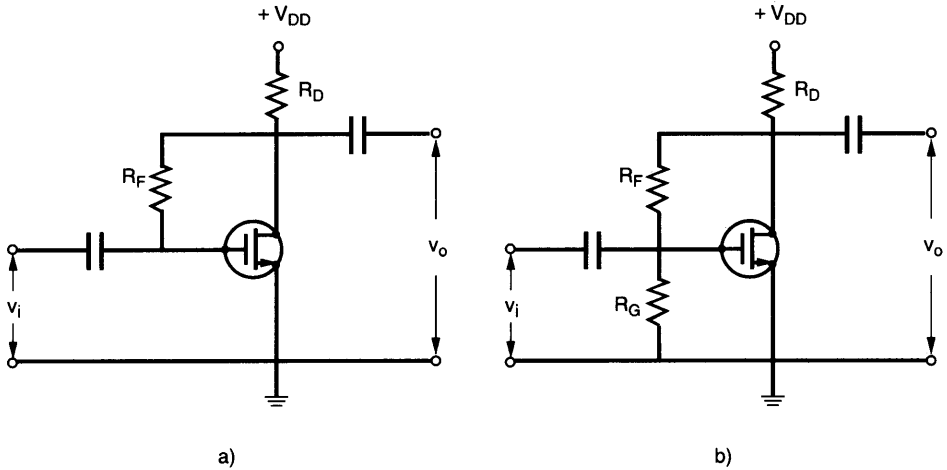


Fig. 9.18. *Diferentes circuitos de polarización utilizados en amplificadores con un transistor MOSFET.*

Dado que en un MOSFET no circula corriente a través de la puerta, la resistencia R_F hace que la puerta quede polarizada al mismo potencial del drenador. En estas circunstancias, el punto de trabajo está situado en un diagrama I_D - V_D en la intersección de la recta de carga con la línea formada por el lugar geométrico de los puntos de las curvas características que cumplen la condición $V_{DS} = V_{GS}$. En la figura 9.19 se han trazado las curvas características del transistor junto con la recta de carga, y en línea de trazos se ha señalado los puntos que cumplen dicha condición. El punto Q_1 corresponde por tanto al punto de funcionamiento del transistor.

Utilizando el circuito de polarización de puerta descrito en la fig. 9.18a puede ocurrir que el punto de funcionamiento quede en el borde de la región de saturación de las curvas características, lo cual no es muy conveniente ya que en esta zona se pierde la linealidad entre la señal de salida en relación con la señal de entrada. A este respecto hay que señalar que el potencial de saturación viene determinado por la relación: $V_{D,sat} = V_{GS} - V_T$, siendo V_T el potencial umbral necesario para producir la inversión de portadores en el transistor (véase el

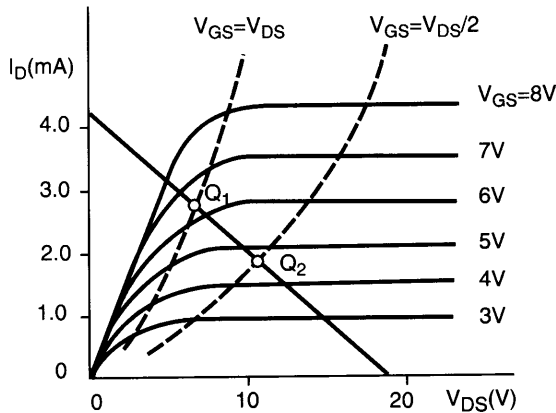


Fig. 9.19. Determinación del punto de trabajo en un amplificador basado en el transistor MOSFET. Las líneas a trazos corresponden a los puntos de funcionamiento que cumplen $V_{GS}=V_{DS}$ y $V_{GS}= V_{DS} / 2$.

apartado 8.6.2). En los transistores de silicio el valor de V_T suele ser de 1 V o menor. Para conseguir que el punto de funcionamiento se sitúe bien en el centro de la región de saturación, frecuentemente se emplea el circuito de polarización de puerta de la fig. 9.18b en el cual se ha incluido una resistencia, R_G , entre el terminal de puerta y tierra. Es fácil demostrar que en este caso el punto de funcionamiento viene dado por la intersección de la recta de carga y la curva que une los puntos que cumplen la ecuación:

$$V_{GS} = \frac{R_G}{R_F + R_G} V_{DS} \tag{9.27}$$

En la fig. 9.19 se ha representado en línea de trazos la curva que cumple la condición anterior para el caso particular $R_F=R_G$, y por tanto $V_{GS} = V_{DS} / 2$. Obsérvese que el nuevo punto de funcionamiento, Q_2 , queda muy por debajo del Q_1 sobre la recta de carga, y por tanto mucho más alejado del borde de la región de saturación de las curvas características.

Una diferencia importante del circuito amplificador del MOSFET respecto a los circuitos con JFET es que la resistencia de polarización está directamente unida al terminal del drenador en lugar de ir al terminal positivo de la fuente de alimentación. La ventaja de este nuevo circuito de polarización de puerta reside en la mayor estabilidad de la corriente de

drenador. En efecto, si por alguna causa externa la corriente I_D tiende a disminuir, la caída de voltaje a través de la resistencia de drenador, R_D , también disminuiría. Esto provocaría un aumento del potencial en el terminal de drenador respecto a tierra, V_{DS} , y por consiguiente del potencial de puerta, V_{GS} . Con ello, la corriente a través del transistor aumentaría, compensando así la disminución de la corriente de drenador.

De toda la discusión precedente se desprende que los MOSFET, al igual que los JFET, pueden ser utilizados en circuitos amplificadores presentando unas buenas características de amplificación, y en particular una alta impedancia de entrada. Dado que en los MOSFET la puerta está separada del resto del semiconductor por una capa de SiO_2 extremadamente aislante, **la resistencia de entrada de un MOSFET es excepcionalmente elevada, con valor prácticamente infinito para todos los efectos**. Este hecho permite además utilizar directamente varios amplificadores en cascada sin necesidad de intercalar condensadores de acoplo de un amplificador a otro, ya que el voltaje de polarización del drenador sobre la que va superpuesta la señal de salida de una etapa puede utilizarse para polarizar la puerta a la entrada de la siguiente etapa. La ausencia de condensadores de acoplo permite simplificar el circuito y mejorar asimismo la respuesta para señales de frecuencia más baja (en el siguiente apartado se discute este efecto con más detalle). El hecho de que no exista corriente de minoritarios a través del circuito de puerta hace también a estos amplificadores menos sensibles a los cambios de temperatura, lo cual constituye una ventaja considerable de los MOSFET.

En cualquier caso, el mayor campo de utilización de los MOSFET de silicio reside en los circuitos integrados, sobre todo en los de tipo digital. Aunque la respuesta en frecuencia de los MOSFET no iguala a la de los bipolares (debido a la capacidad asociada al aislante de puerta), sin embargo ofrecen características muy adecuadas para la integración a gran escala (pequeño tamaño, menor número de etapas de fabricación, menor consumo en circuitos de conmutación, etc.). Algunos de estos aspectos serán discutidos más adelante en relación con los procesos de fabricación de los circuitos integrados (cap. XIII).

9.7. RESPUESTA EN FRECUENCIA DE LOS AMPLIFICADORES

En general, los amplificadores descritos en este capítulo están diseñados para aumentar el voltaje de una señal sinusoidal cuya frecuencia esté en un rango que va desde unas decenas de hercios hasta las decenas o centenas de kilohercios (rango denominado de *baja frecuencia* o también de *audiofrecuencia*). Fuera de este rango la señal de salida puede quedar atenuada debido a que el factor de amplificación es más bajo. Este hecho tiene importancia ya que da lugar a que señales no sinusoidales, que contengan armónicos de frecuencias elevadas, puedan resultar distorsionadas en el amplificador. Se hace, pues, conveniente conocer los límites de frecuencia dentro de los cuales el factor de amplificación se mantiene constante (esto es lo que se conoce como *anchura de banda de un amplificador*) así como los factores que determinan estos límites.

En la fig. 9.20 se muestra la variación con la frecuencia de la señal de entrada del factor de amplificación normalizado, A/A_0 , de un circuito amplificador típico, bien sea con transistores bipolares o de efecto campo trabajando en la configuración de emisor o de fuente común. Según se aprecia, la curva de respuesta en frecuencia presenta en todos estos casos un tramo horizontal, en el cual A se mantiene constante en un valor fijo: $A = A_0$. El tramo horizontal está limitado por dos frecuencias extremas, por encima o por debajo de las cuales la ganancia del amplificador decae notablemente por debajo del valor a frecuencias medias. Veamos las causas que determinan este efecto.

9.7.1. Región de bajas frecuencias

Como es sabido, cualquier elemento de un circuito con impedancia compleja (condensadores o inducciones) presenta una reactancia (capacitiva o inductiva) cuyo valor depende de la frecuencia. Consideremos por ejemplo el circuito de emisor común presentado en la fig. 9.10. Según vimos más arriba, la capacitancia (o reactancia capacitiva) del condensador de acoplo, C_i , situado a la entrada del amplificador viene dada por: $X_c = 1/2\pi f C_i$, siendo f la frecuencia de la señal de entrada. Normalmente, el condensador C_i se elige de forma que su impedancia a las frecuencias de trabajo sea "suficientemente baja" con objeto de que la señal pase a través del condensador sin sufrir atenuación. Es evidente que, si disminuimos la frecuencia de la señal, eventualmente se alcanzará un valor de X_c para el cual la señal de entrada quede sensiblemente atenuada.

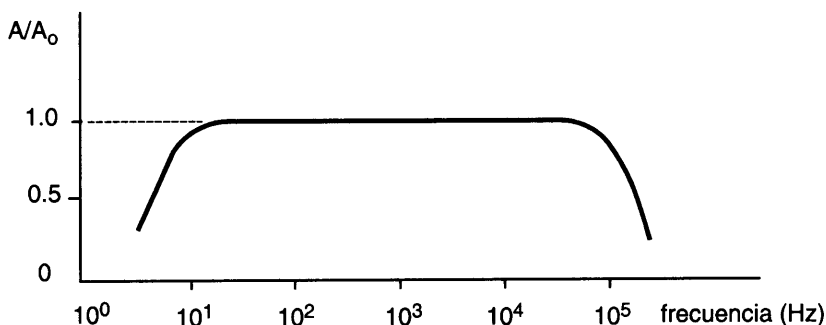


Fig. 9.20. Curva de variación del factor de amplificación (normalizado) con la frecuencia de un circuito amplificador.

El efecto de atenuación de la señal de entrada por el condensador C_i puede discutirse con ayuda de la fig. 9.21a, en la cual se presenta un esquema de un circuito RC equivalente (obtenido mediante el teorema de Thévenin) correspondiente a la entrada del amplificador (*filtro pasa-alta*). La resistencia R_i representa la resistencia dinámica de entrada del amplificador de emisor común en el punto de operación ($R_i = r'$) y es, según vimos, la que recibe la señal, v_{be} , entre los terminales de emisor y base. Asimismo, el condensador corresponde al condensador de acoplo para la señal de entrada ($C_1 = C_i$). Suponiendo que el condensador es un elemento puramente capacitivo, el conjunto resistencia-condensador de la fig. 9.21a puede considerarse como un divisor de tensión para la señal de entrada, v_i . Un análisis del comportamiento de este circuito en corriente alterna, utilizando variable compleja, permite calcular fácilmente el cociente entre el voltaje v_{be} que realmente se introduce en el terminal de base del amplificador, y la señal de entrada, v_i . El cociente $v_{be}/v_i = A_{if}$ representa el *factor de atenuación* de voltaje a baja frecuencia para el circuito de entrada del amplificador, y viene dado por:

$$A_{if} = \frac{v_{be}}{v_i} = \frac{R_i}{R_i - jX_{c1}} = \frac{1}{1 - j(f_{c1}/f)} \quad [9.28]$$

donde $j = \sqrt{-1}$, $X_{c1} = 1/2\pi fC_1$ y f_{c1} viene dado por:

$$f_{c1} = \frac{1}{2\pi R_i C_1} \quad [9.29]$$

Por tanto, el módulo $|A_{if}|$, y el ángulo de fase, θ_{if} (adelanto), se pueden expresar como:

$$|A_{if}| = \frac{1}{[1 + (f_{c1}/f)^2]^{1/2}} \quad [9.30]$$

$$\theta_{if} = \arctg \frac{f_{c1}}{f} \quad [9.31]$$

A frecuencias altas $X_{c1} \rightarrow 0$, por lo que la fracción v_{be}/v_i dada por la ec. [9.28] tiende a la unidad, es decir la señal entra en el amplificador sin sufrir atenuación. Al mismo tiempo el ángulo de fase es cero. En cambio, si se disminuye la frecuencia de la señal el valor de X_{c1} aumenta y la fracción de señal que entra en el amplificador disminuye (fig. 9.21b). Para una frecuencia $f = f_{c1}$ la señal de entrada queda atenuada en módulo en el factor de $1/\sqrt{2} = 0.707$. De las ecuaciones anteriores se deduce que a esta frecuencia, denominada *frecuencia de corte*, se cumple la igualdad $X_{c1} = R_i$, esto, es la reactancia del condensador de acoplo se iguala a la resistencia entre la base y el emisor.

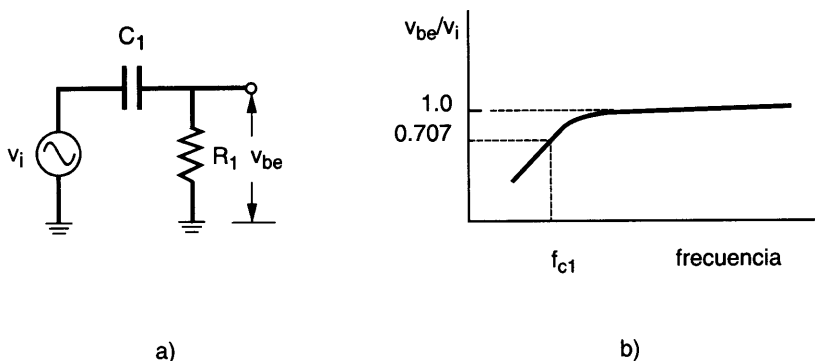


Fig. 9.21. a) Circuito equivalente de Thévenin para la entrada de un amplificador de emisor común provisto de un condensador de acoplamiento, C_1 . b) Curva de variación con la frecuencia del cociente v_{be}/v_i mostrando la frecuencia de corte, f_{c1} .

Un análisis similar para la corriente, permite concluir que la corriente de entrada en el circuito RC sufre un efecto similar, atenuándose también en el factor $1/\sqrt{2}$ a la frecuencia de corte. De este modo, **la atenuación de la potencia de señal de entrada en el circuito de acoplamiento del amplificador sería para la frecuencia de corte igual a $1/2$.**

El efecto del condensador de acoplo a la salida del amplificador, C_o (fig. 9.10), es muy similar al de entrada, por lo que este condensador dará lugar a una atenuación de $1/\sqrt{2}$ en la señal de salida cuando la frecuencia disminuye hasta una cierta frecuencia de corte. Del mismo modo, el condensador de paso C_E en la resistencia de emisor (fig. 9.15) también contribuye a la atenuación del factor de amplificación del amplificador al bajar la frecuencia, ya que su misión es, según se ha señalado, disminuir la resistencia R_E del amplificador y evitar un efecto de realimentación de la señal de salida. De toda esta discusión se puede concluir que cada uno de los condensadores del circuito tiene una frecuencia de corte dada por una ecuación similar a la ec. [9.29]. De todas ellas, la más alta será la que origina la caída de la ganancia del amplificador en el factor 0.707 en el lado de bajas frecuencias, tal como se especifica en la fig. 9.20, y es la que se denomina *frecuencia de corte inferior* del circuito amplificador.

Para conseguir que la frecuencia de corte inferior sea lo más baja posible y aumentar la anchura de banda del amplificador se utilizan condensadores de acoplo de capacidad muy alta, típicamente unos 1-10 μF . Si la resistencia r' de entrada es por ejemplo de unos 2000

ohmios, valor típico, la frecuencia de corte inferior se situaría entonces en unos 10-100 Hz. Obviamente, si se eliminan los condensadores de acoplo y el de paso en el emisor mejora notablemente la respuesta en frecuencia, siendo posible amplificar señales de frecuencia cero, es decir, en continua o con variación muy lenta. Esto realmente ocurre en cierto tipo de amplificadores en los cuales la mejora en el lado de baja frecuencia se hace a expensas de una pérdida de ganancia a frecuencias intermedias (debido a que R_E ya no queda cortocircuitada en este rango), con pérdida incluso de estabilidad. Este nuevo tipo de amplificadores, denominados de *acoplamiento directo*, juegan un papel muy importante en electrónica y su comportamiento será descrito en el siguiente capítulo.

9.7.2. Región de frecuencias altas

En la región de altas frecuencias, la curva de la ganancia disminuye de nuevo a partir de una cierta frecuencia crítica. Diversos efectos pueden contribuir a esta nueva caída del factor de amplificación. Entre ellos queremos destacar: i) la disminución del factor α (ó β) del transistor, y ii) la presencia de *capacidades e inducciones parásitas* en el circuito, tales como la capacidad asociada a las uniones de emisor y colector, así como las capacidades e inducciones debidas al cableado y a otros componentes presentes en el amplificador.

El descenso del factor α está relacionado con el tiempo medio de tránsito de los portadores a través de la base en su paso desde el emisor hacia el colector cuando el transistor opera en la región activa. Si el tiempo de tránsito es mayor que el período de la señal de entrada, los portadores (huecos, en el caso de un transistor pnp) no tienen tiempo suficiente para responder a las variaciones de la señal de entrada y una fracción de ellos queda acumulada en la región de base, aumentando con ello la tasa de recombinación en la base. La corriente en el colector será entonces menor que la obtenida en el rango intermedio de frecuencias, lo que da lugar a una reducción del factor de ganancia en corriente alterna, α . El factor β , así mismo, se verá afectado en una proporción mayor (recuérdese que $\beta = \alpha / (1 - \alpha)$, ec. 6.14), por lo que en conjunto el factor de amplificación también quedará mermado. El tiempo de tránsito a través de la base depende sobre todo de la anchura de la base y de la movilidad de los portadores. Evidentemente, cuanto menor sea el tiempo de tránsito más alta será la frecuencia de corte del transistor debida a este efecto.

El efecto de las capacidades parásitas en el circuito puede discutirse con ayuda del circuito mostrado en la fig. 9.22a, en el que se representa en línea de trazos las capacidades asociadas a las uniones de emisor y colector en un circuito amplificador de emisor común. Según vimos en el capítulo 6, las uniones de emisor y colector están polarizadas, respectivamente, en directo e inverso cuando el transistor trabaja en la región activa, y por tanto el comportamiento de la región de carga espacial se puede comparar al de un condensador. Estas capacidades, cuyo valor en conjunto no excede unos pocos picofaradios, no ejercen ningún efecto apreciable en la región de frecuencias intermedias o bajas ya que su reactancia, X_{c2} , (en paralelo con la señal aplicada al transistor) es alta. En cambio a altas frecuencias X_{c2} disminu-

ye, llegando incluso a cortocircuitar la señal en los terminales de salida del transistor. En estas condiciones, la señal de salida del amplificador puede verse seriamente afectada, llegando a alcanzar valores muy bajos próximos a cero.

El análisis de la influencia de estas capacidades parásitas es algo más complejo que en el caso de bajas frecuencias, ya que es preciso considerar también las capacidades e inducciones debidas a los componentes e interconexiones que se hallan distribuidos por todo el circuito amplificador, por lo que sólo haremos una descripción cualitativa. Para hacer este análisis es conveniente recurrir al circuito RC equivalente del amplificador visto entre los terminales de

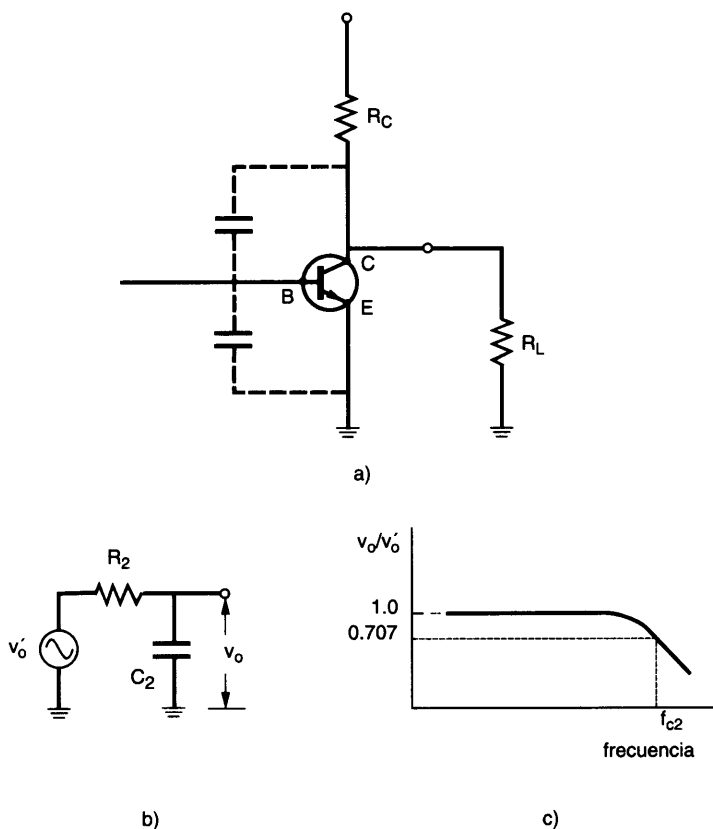


Fig. 9.22. a) Representación de las capacidades asociadas a las uniones de emisor y colector de un circuito amplificador de emisor común (línea a trazos). b) Circuito equivalente de Thévenin visto entre los terminales de salida del amplificador. c) Curva de variación con la frecuencia del cociente v_o/v_o' , mostrando la frecuencia de corte, f_{c2} .

salida, mostrado de forma cualitativa en la fig. 9.22b (*filtro pasa-baja*). En este circuito v'_o representa la señal que existiría entre los terminales de colector y emisor si no existiera la influencia de las capacidades parásitas, R_2 la resistencia serie equivalente a la resistencia de colector y de carga así como otras resistencias asociadas al circuito y C_2 la capacidad equivalente de las uniones de emisor y colector y otras capacidades parásitas. A partir del circuito de la fig. 9.22b es fácil demostrar que la relación entre la señal de salida del amplificador, v_o , y el de la señal v'_o viene dada por:

$$A_{hf} = \frac{v_o}{v'_o} = \frac{-jX_{c2}}{R_2 - jX_{c2}} = \frac{1}{1 + j(f/f_{c2})} \quad [9.32]$$

En esta ecuación A_{hf} es el factor de atenuación de la ganancia a altas frecuencias, $X_{c2} = 1/2\pi f C_2$ y f_{c2} una frecuencia dada por una expresión similar a la de bajas frecuencias, esto es:

$$f_{c2} = \frac{1}{2\pi R_2 C_2} \quad [9.33]$$

El módulo, $|A_{hf}|$, y el ángulo de fase, θ_{hf} (retraso), de este factor vienen dados por:

$$|A_{hf}| = \frac{1}{[1 + (f/f_{c2})^2]^{1/2}} \quad [9.34]$$

$$\theta_{hf} = - \arctan \frac{f}{f_{c2}} \quad [9.35]$$

A frecuencias bajas e intermedias X_{c2} toma un valor alto, siendo $X_{c2} \gg R_2$, por lo que el factor A_{hf} tiende a la unidad y el ángulo de fase es próximo a cero. En cambio a frecuencias altas X_{c2} disminuye con la frecuencia y el factor A_{hf} tiende a cero (fig. 9.22c). En esta región de altas frecuencias, existe una frecuencia, $f = f_{c2}$, en la que se cumple $X_{c2} = R_2$. Para esta frecuencia, denominada *de corte superior*, la señal de salida queda atenuada en el factor $1/\sqrt{2} = 0.707$. Para valores típicos de $C_2 = 1\text{-}10$ pF, y $R_2 = 1$ Mohm, el valor de f_{c2} resultante es $f_{c2} = 10\text{-}100$ KHz.

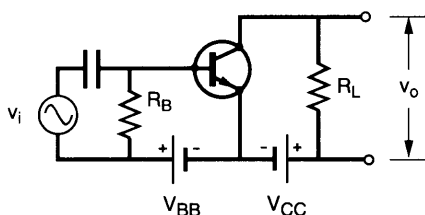
De los dos efectos mencionados de reducción de la señal de salida (disminución del factor α y capacidades parásitas), aquel que tenga la frecuencia más baja de corte será el que finalmente determine la frecuencia de corte superior del circuito amplificador. De la discusión se desprende que para diseñar circuitos amplificadores que trabajen a frecuencias muy altas, es preciso evitar al máximo las capacidades parásitas que aparecen en paralelo con la

señal y disminuir al máximo el tiempo de tránsito de los portadores en la región de base (o en la del canal cuando se trata de los FET). La utilización de transistores del tipo MESFET de GaAs en circuitos integrados ha permitido la preparación de amplificadores con un rango útil de operación que llega hasta las frecuencias de microondas (véase apartado 8.4).

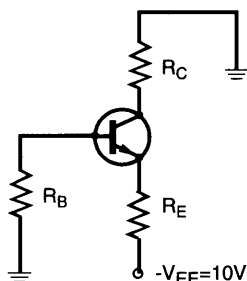
Podemos definir ahora de forma cuantitativa la anchura de banda de un amplificador, Δf , como la región de frecuencias que va desde f_{c1} hasta f_{c2} , es decir, $\Delta f = f_{c2} - f_{c1}$. En esta región de frecuencias el factor de amplificación es prácticamente constante, o en su caso se atenúa en un factor menor que 0.707. Esto quiere decir que cualquier señal, cuyas componentes armónicas con amplitud apreciable caen dentro de este rango, pasará a través del amplificador sin excesiva distorsión.

CUESTIONES Y PROBLEMAS

9.1 Hallar el punto de funcionamiento y el factor de amplificación en el circuito de la figura, con $V_{CC} = 20$ V, $V_{BB} = 2.0$ V, $R_L = 2600$ ohm, $R_B = 20K$ ohm y $\beta = 50$ (suponer un transistor de silicio cuya curva I-V de entrada viene dada en la fig. 9.5, con $V_{BE} = 0.6$ V). Hacer un diagrama de las señales de entrada y salida.

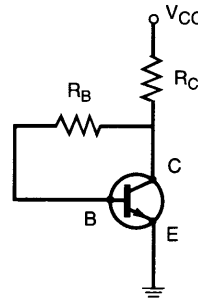


9.2 En el circuito de la figura se tiene: $R_C = R_E = 1K$ ohm, $R_B = 270K$ ohm, $V_{EE} = -10$ V y $\beta = 100$. Determinar: a) la región de operación del transistor, y b) los valores de I_B , I_C y V_{CE} (suponer $V_{BE} = 0.6$ V).



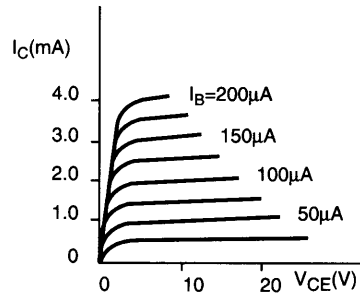
9.3 Para el circuito de la fig. 9.13, con un transistor de silicio ($V_{BE} = 0.6$ V) y $\beta = 60$, hallar el punto de funcionamiento suponiendo $V_{CC} = 12$ V, $R_1 = 140K$ ohm, $R_2 = 25K$ ohm, $R_C = 5$ K ohm y $R_E = 1K$ ohm.

- 9.4** En el circuito de la figura se tiene $V_{CC} = 9V$. Si el transistor posee un factor de ganancia en corriente $\beta = 99$, a) calcular R_C y R_B de forma que $I_C = 5$ mA y $V_{CE} = 5V$, b) suponiendo que el valor de β fuera 49, ¿cuál sería entonces el nuevo valor de I_C y de V_{CE} ? (tómese $V_{BE} = 0.6$ V).

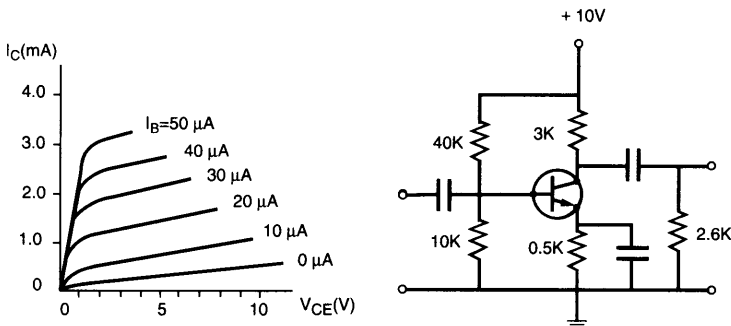


- 9.5** En un transistor de silicio tipo npn conectado en la configuración de emisor común con autopolarización de base, con $V_{CC} = 24$ V, $R_C = 6K$ ohm y $\beta = 60$, se desea que el punto de funcionamiento esté localizado en $V_{CE} = 12$ V e $I_C = 1.6$ mA y que el factor de estabilidad térmico, dado por la ec. [9.19], sea 3 como máximo. Calcular los valores de las resistencias del circuito de autopolarización (tómese $V_{BE} = 0.6$ V).

- 9.6** Un transistor de silicio, cuyas características vienen dadas en la figura, se utiliza en un circuito amplificador similar al de la fig. 9.13a, con $V_{CC} = 20$ V, $R_C = 5K$ ohm, $R_1 = 100K$ ohm, $R_2 = 10K$ ohm y $R_E = 1K$ ohm. Suponiendo que $V_{BE} = 0.6$ V, calcular el punto de funcionamiento.

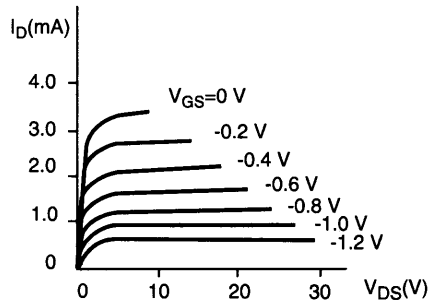


- 9.7** El transistor del circuito amplificador adjunto tiene las curvas características de la figura. Determinar el punto de funcionamiento y calcular la corriente a través de la resistencia de carga cuando se introduce en la base una señal alterna de $10 \mu A$ de valor de pico. ¿Cuánto vale el factor de amplificación en corriente de este circuito?

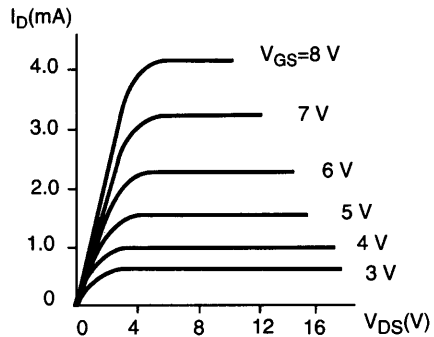
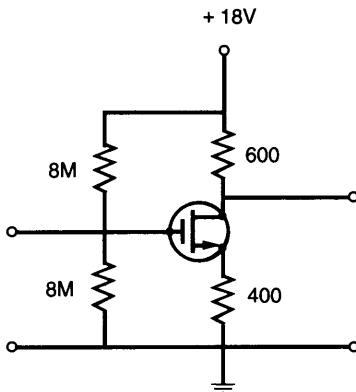


9.8 En un circuito amplificador formado por un transistor de silicio tipo npn, con autopolarización de base, se tiene: $R_1 + R_2 = 25\text{K ohm}$, $R_C = 1\text{K ohm}$ y $V_{CC} = 10\text{ V}$. Suponiendo $\beta = 49$, determinar los valores de R_1 , R_2 y R_E de forma que el punto de funcionamiento del transistor esté situado en $V_{CE} = 4.1\text{ V}$ e $I_C = 4.9\text{ mA}$.

9.9 Un FET que tiene las curvas características de la figura se utiliza en el circuito amplificador de la fig. 9.16. Determinar el punto de funcionamiento y obtener el factor de amplificación en voltaje cuando la señal de entrada tiene un valor de pico de 0.2 V , suponiendo que $V_{DD} = 25\text{ V}$, $R_D = 10\text{K ohm}$, $R_1 = 600\text{K ohm}$, $R_2 = 24\text{K ohm}$ y $R_S = 2\text{K ohm}$.



9.10 Determinar el punto de funcionamiento en el circuito amplificador que se muestra en la figura adjunta, utilizando las curvas características de la figura para el MOSFET.



9.11 En el circuito amplificador de la fig. 9.16 se tiene: $R_1 = 20\text{M ohm}$, $R_2 = 1\text{M ohm}$, $R_D = 7\text{K ohm}$, $R_S = 3\text{K ohm}$ y $V_{DD} = 20\text{ V}$. Localizar el punto de funcionamiento y hallar el factor de amplificación en voltaje para una señal alterna con valor de pico de 0.25 V , suponiendo que las curvas características del transistor son las de la fig. 9.17.

CAPITULO XI

AMPLIFICADORES REALIMENTADOS Y OPERACIONALES

En los amplificadores realimentados una fracción de la señal de salida se vuelve a introducir a la entrada, mejorando de este modo muchas de las características de los amplificadores. Cuando esta fracción tiende a disminuir la señal de entrada entonces la realimentación se denomina negativa, siendo ésta la empleada más usualmente en circuitos amplificadores. En estas condiciones, el factor de amplificación del dispositivo se ve algo disminuido, aunque mejora mucho tanto la respuesta con la frecuencia como la estabilidad del amplificador. Las resistencias de entrada y de salida también se acercan más a las del amplificador ideal. Una clase especial de realimentación es la denominada realimentación operacional que constituye la base de los amplificadores operacionales. Estos amplificadores fueron inicialmente propuestos para la realización de operaciones matemáticas en las computadoras analógicas. Sin embargo, hoy día se utilizan prácticamente en todas las áreas de aplicación de la electrónica analógica, tales como fuentes de alimentación, generadores de funciones, instrumentos electrónicos de medida, etc. Debido a la importancia de estos amplificadores, en este capítulo se pretende dar una visión general de los fundamentos de la realimentación para estudiar después los amplificadores operacionales.

11.1. AMPLIFICADORES REALIMENTADOS

Según vimos en el capítulo anterior, el factor de amplificación de voltaje, A , de un amplificador ordinario (representado esquemáticamente en la fig. 11.1a mediante un triángulo) viene dado por la relación entre el voltaje de la señal de salida, v_o , y el voltaje de la señal

de entrada, v_i , es decir:

$$A = v_o / v_i$$

A este cociente lo denominaremos *factor de amplificación en circuito abierto*. Cuando se establece un lazo de realimentación en el amplificador anterior, como el indicado en la fig. 11.1b, la señal de salida, que ahora representaremos por v'_o , es tomada en el *punto de muestreo M*, y una fracción de esta señal, v_f , dada por $v_f = \beta v'_o$ (β es el denominado *factor de realimentación*) es introducida de nuevo a la entrada en el punto sumador, S. La señal total de entrada, v'_i , en el amplificador A, es ahora igual a la señal original, v_i , más la señal de realimentación, v_f , esto es:

$$v'_i = v_i + v_f = v_i + \beta v'_o \quad [11.1]$$

Así pues, la señal de salida del amplificador cumplirá la ecuación:

$$v'_o = Av'_i = A(v_i + \beta v'_o)$$

la cual se puede escribir en la forma:

$$Av_i = (1 - \beta A) v'_o$$

Si definimos $A_f = v'_o / v_i$ como el factor de amplificación del circuito con realimentación se tendrá que:

$$A_f = \frac{A}{1 - \beta A} \quad [11.2]$$

Es decir, el factor de amplificación del amplificador realimentado puede ser mayor o menor que el del correspondiente amplificador simple dependiendo del signo del producto βA . En el caso más general, esto es, cuando existe un cambio de fase a la salida del amplificador o del circuito de realimentación, los valores de A y β son complejos. En lo que sigue, y con objeto de simplificar la discusión, supondremos que tanto A como β son siempre números reales, positivos o negativos. Esto implica que la señal de salida del amplificador y del circuito de realimentación está en fase o en oposición de fase con la señal de entrada, según el caso.

El caso más interesante se produce cuando el factor de realimentación β es negativo, o lo que es lo mismo cuando el voltaje $v_f = \beta v'_o$ que se realimenta desde la salida a la entrada del amplificador presenta una diferencia de fase de 180° con respecto al voltaje de la señal de entrada, v_i (*realimentación negativa*). Al ser β negativo, la ganancia del amplificador

realimentado, A_f , disminuye en relación al circuito sin realimentar, ya que queda dividida por el factor $1-\beta A$ (que es mayor que la unidad). Para hacerse una idea de lo que puede disminuir el factor de amplificación cuando el circuito se cierra mediante un lazo de realimentación negativa, en la Tabla 11.1 se presentan los valores calculados del cociente A_f/A y del factor A_f correspondientes a distintos valores de β , para el caso de que A valga 10. Obsérvese, por ejemplo, que cuando el factor de realimentación β es tan sólo el 10 % (es decir, una décima parte de la señal de salida es introducida en oposición de fase a la entrada), la ganancia del amplificador realimentado se reduce ya a la mitad.

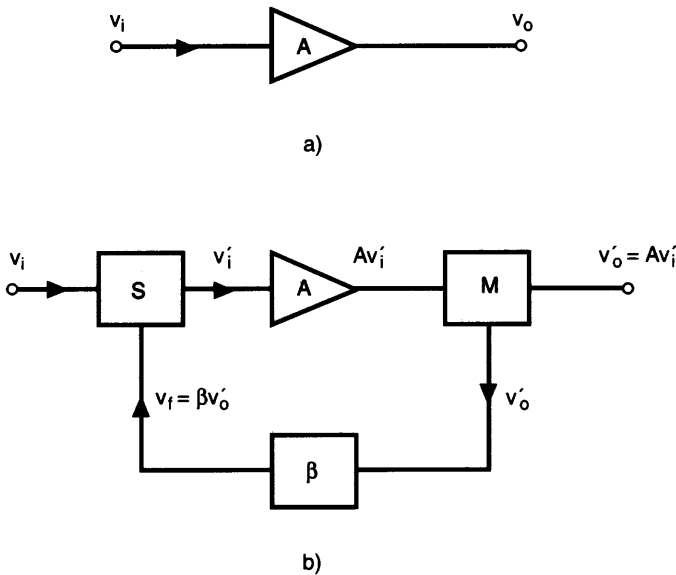


Fig. 11.1. a) Diagrama de bloques de un circuito amplificador simple. b) Diagrama de bloques del amplificador anterior con realimentación para señal de voltaje.

De la ec. [11.2] se deduce que si β es positivo (*realimentación positiva*) el factor de amplificación A_f aumenta con respecto al caso en que no hay realimentación. Sin embargo, la realimentación positiva tiene el inconveniente de que aumenta también la distorsión de la señal amplificada y reduce asimismo la estabilidad del dispositivo. Por esta razón esta disposición raramente se utiliza en el diseño de amplificadores.

De todos modos, hay un caso especial de realimentación positiva que tiene mucho interés, y es aquél en el que el factor βA se iguala a la unidad, esto es $\beta A = 1$. De acuerdo con la ec. [11.2], el factor de amplificación A_f se hace entonces infinito. En estas circunstancias,

incluso en ausencia de señal a la entrada (en realidad siempre existe una señal de ruido) el dispositivo sería capaz de producir una señal de salida. Podría decirse que este caso es el de máxima inestabilidad del amplificador, aunque es posible sacar ventaja de este efecto para producir un circuito oscilador. Efectivamente, un *circuito oscilador* no es más que un dispositivo amplificador, capaz de suministrar una señal de salida, de una frecuencia determinada, sin necesidad de introducir una señal a la entrada. Para conseguir esto, la condición $\beta A = 1$ se ha de cumplir solamente para las frecuencias de interés, por lo que el circuito debe disponer de un sistema de filtraje o sintonía que elimine el resto de las frecuencias no deseadas en la salida. Los osciladores constituyen la base de un gran número de dispositivos generadores de señal (sinusoidal, cuadrada, etc.), los cuales encuentran un uso muy extendido en los circuitos electrónicos. Dado el carácter introductorio de este libro, no se hará ninguna descripción de estos dispositivos. El lector interesado puede encontrar una información detallada en textos más avanzados.

TABLA 11.1

VARIACION DEL COCIENTE A_f/A Y DEL FACTOR DE AMPLIFICACION A_f EN FUNCION DE LA FRACCION DE REALIMENTACION NEGATIVA, β (para $A = 10$)

Factor de realimentación, β (%)	Cociente $A_f/A = 1 / (1-\beta A)$	Factor de amplificación, A_f
0 %	1.00	10.0
1 %	0.91	9.1
2 %	0.83	8.3
10 %	0.50	5.0
20 %	0.33	3.3
30 %	0.25	2.5
50 %	0.17	1.7
100 %	0.09	0.9

Si en la ecuación [11.2] dividimos el numerador y denominador por A se obtiene:

$$A_f = \frac{1}{(1 / A) - \beta}$$

En el caso particular de $\beta \gg 1/A$ resulta:

$$A_f = -1/\beta \quad [11.3]$$

Esta condición se presenta frecuentemente cuando la ganancia A del amplificador no realimentado es muy elevada, como ocurre por ejemplo en los amplificadores operacionales, estudiados más adelante. En estos amplificadores el factor A puede llegar hasta 10^5 . Se tiene en estos casos que la amplificación del circuito realimentado, A_f , depende tan sólo de β , esto es, de las características del lazo de realimentación. El factor de amplificación es entonces independiente de las desviaciones del comportamiento ideal de los transistores, así como también de posibles variaciones en los componentes e incluso de las originadas por las fuentes de alimentación del circuito amplificador. Como veremos más adelante, este es uno de los aspectos más relevantes de los amplificadores operacionales cuando trabajan con un lazo de realimentación negativa.

11.2. CARACTERÍSTICAS DE LOS AMPLIFICADORES REALIMENTADOS

11.2.1. Distintas configuraciones del circuito de realimentación

Según sea la disposición de las redes de muestreo y sumadora, representadas por M y S en la fig. 11.1b, se puede tener cuatro configuraciones posibles de amplificadores realimentados, representados en fig. 11.2. En las figs. 11.2a y 11.2b la toma de la señal para la realimentación se hace en paralelo con la señal de salida, y se dice entonces que se muestrea la tensión de salida. En cambio, en las figs. 11.2c y 11.2d la red de realimentación está conectada en serie con la salida, diciéndose en este caso que se muestrea la corriente de salida.

Similarmente, la red sumadora puede conectar la señal de realimentación bien con entrada en serie (figs. 11.2a y 11.2c) o con entrada en paralelo (figs. 11.2b y 11.2d). La red de realimentación, representada en la figura por el bloque β , puede contener en el caso más general un circuito complejo formado por un conjunto de componentes activos y pasivos aunque frecuentemente la red se reduce simplemente a una serie de resistencias. A menudo se hace difícil discernir en un circuito el tipo de realimentación, serie o paralelo, en un amplificador e incluso conocer la ganancia en circuito abierto. En la sección 11.3 se presentan algunos ejemplos de circuitos típicos de realimentación.

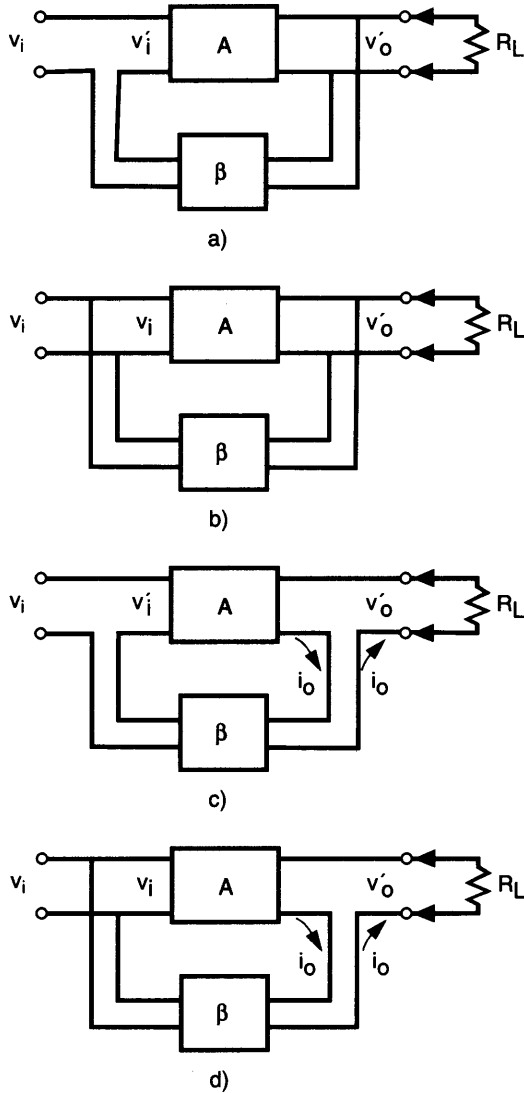


Fig. 11.2. Diferentes configuraciones de las redes de muestreo y sumadora: a) Realimentación de voltaje con entrada serie. b) Realimentación de voltaje con entrada paralelo. c) Realimentación de corriente con entrada serie, y d) Realimentación de corriente con entrada paralelo.

11.2.2. Estabilidad en la amplificación

En el capítulo noveno vimos que las variaciones de los transistores debidas a cambios de temperatura, cambios en los parámetros de los transistores, etc, afectan la estabilidad de la amplificación. Sin embargo, estos problemas son mucho menores en los amplificadores realimentados, tal como hemos mencionado. En efecto, de la ecuación [11.2] se deduce que un cambio ΔA en el factor de amplificación produce un cambio ΔA_f dado por:

$$\Delta A_f = \frac{\Delta A}{(1 - \beta A)^2}$$

Dividiendo esta ecuación miembro a miembro por la ecuación [11.2] se tiene:

$$\frac{\Delta A_f}{A_f} = \frac{1}{1 - \beta A} \frac{\Delta A}{A} \quad [11.4]$$

De esta ecuación se desprende que en realimentación negativa ($\beta < 0$) la variación relativa de A_f es menor que la de A en un factor igual al cociente $1/(1-\beta A)$. De acuerdo con los datos de la tabla 11.1, cuando el factor de realimentación negativa es, por ejemplo, del 10 % ($\beta = 0.1$) y el factor de amplificación $A=10$, la relación $1/(1-\beta A)$ vale 0.5. Así pues, la estabilidad del circuito aumenta en este caso en un factor 2. Se concluye, por tanto, que la estabilidad del amplificador con realimentación negativa puede ser muy superior a la del amplificador en lazo abierto.

11.2.3. Resistencia de entrada y de salida del amplificador realimentado

Consideremos un amplificador de voltaje sin realimentación como un circuito de dos puertas, tal como el que ha sido descrito anteriormente en la fig. 9.1b, en el cual se conecta en la puerta de entrada un generador de señal, v_s , de resistencia serie R_s , y en la de salida una resistencia de carga, R_L (fig. 11.3a). Las resistencias de entrada y de salida del amplificador vienen representadas por R_i y R_o , respectivamente. El amplificador de voltaje ideal debe proporcionar un voltaje de salida v_o proporcional a la señal del generador de entrada v_s . Para que esto ocurra se ha de verificar que R_o tenga un valor muy pequeño frente a la resistencia de carga R_L . Al mismo tiempo, R_i ha de ser elevada comparada con la resistencia serie del generador, R_s . En efecto, si $R_o \ll R_L$ se tiene que $v_o \approx A v_i$. Análogamente, si $R_i \gg R_s$, se cumple $v_i \approx v_s$. En estas circunstancias, $v_o \approx A v_i \approx A v_s$. Así pues, según se discutió en el

capítulo noveno, el amplificador ideal de voltaje debe tener una resistencia de entrada prácticamente infinita y una resistencia de salida muy próxima a cero.

Supongamos el caso frecuente de un circuito de realimentación negativa de voltaje y entrada serie, tal como el mostrado en la fig. 11.3b, formado por un amplificador A ordinario, y la red de realimentación β . Este circuito de realimentación se corresponde con el caso señalado anteriormente en la fig. 11.2a. La red de realimentación, β , está formada por una resistencia con una toma intermedia. Esta toma intermedia actúa a modo de divisor de tensión e introduce un voltaje de realimentación, v_f , dado por $v_f = \beta v_o$ (nótese que el parámetro β depende no sólo de las características del divisor de tensión sino también del circuito de entrada). Si la señal es realimentada en serie, como es el caso de la fig. 11.3b, es natural que aumente la resistencia de entrada. En efecto, podemos definir la resistencia de entrada del amplificador realimentado, R_{if} , a través del cociente:

$$R_{if} = v_i / i_i$$

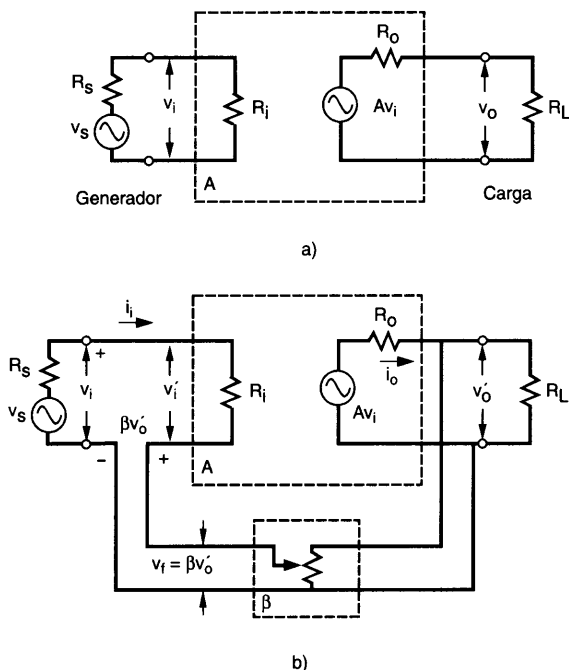


Fig. 11.3. a) Esquema del circuito equivalente de un amplificador de voltaje, según el teorema de Thévenin. b) Esquema del mismo circuito con realimentación de voltaje

esto es, el voltaje de la señal de entrada, v_i , dividido por la corriente que entra en el circuito amplificador, i_i . Tenemos entonces, a partir de la ec. [11.1]:

$$R_{if} = \frac{v'_i - v_f}{i_i} = \frac{v'_i - \beta v'_o}{i_i} = \frac{v'_i}{i_i} (1 - \beta A)$$

Dado que el cociente v'_i / i_i representa la resistencia de entrada del amplificador sin realimentación, esto es $R_i = v'_i / i_i$, se tiene para el amplificador realimentado:

$$R_{if} = R_i (1 - \beta A) \quad [11.5]$$

Así pues, en la realimentación negativa ($\beta < 0$), la resistencia de entrada aumenta en el factor $(1 - \beta A)$ respecto de la que se tiene en el circuito sin realimentación.

Cualitativamente se aprecia también que, al conectar la red de realimentación en paralelo con la salida, deberá disminuir la resistencia de salida, R_{of} . Efectivamente, para obtener R_{of} consideremos los terminales de salida con R_L desconectada y con un voltaje v'_o en la salida. Se define entonces la resistencia de salida a través del cociente (con el voltaje a la entrada $v_i = 0$):

$$R_{of} = v'_o / i_o$$

siendo i_o la corriente en el circuito de salida (sin conectar R_L). Ahora bien, si $v_i = 0$ entonces $v'_i = v_f = \beta v'_o$ por lo que la corriente en el circuito de salida será:

$$i_o = \frac{v'_o - A v'_i}{R_o} = \frac{v'_o - A \beta v'_o}{R_o}$$

y para la impedancia de salida del amplificador realimentado se tendrá:

$$R_{of} = \frac{v'_o}{i_o} = \frac{R_o}{1 - \beta A} \quad [11.6]$$

Es decir, R_{of} disminuye respecto R_o en el factor $1/(1 - \beta A)$. De las expresiones [11.5] y [11.6] se aprecia que el amplificador con realimentación negativa se asimila mucho más al amplificador de voltaje ideal que el amplificador sin realimentación ya que, como hemos visto, su resistencia de entrada es más elevada y la resistencia de salida más pequeña.

11.2.4. Efecto de la realimentación en la anchura de banda

Hemos visto anteriormente (apartado 9.7) que el factor de amplificación de un amplificador disminuye en las regiones extremas de un cierto rango de frecuencias, debido a la atenuación de la señal como consecuencia del efecto producido por los condensadores de acoplo, las capacitancias parásitas en el circuito, etc. Es interesante calcular ahora cómo afectan estos elementos el comportamiento de un circuito realimentado.

En la región de altas frecuencias, podemos expresar el valor del factor de amplificación, A , de un amplificador sin realimentación a una cierta frecuencia, f , utilizando la ec. [9.32] que da la atenuación del factor de ganancia en esta región de altas frecuencias:

$$A = A_o A_{hf} = \frac{A_o}{1 + j(f/f_{c2})} \quad [11.7]$$

siendo A_o la ganancia del amplificador a frecuencias intermedias y A_{hf} el factor de atenuación de la señal en la región de altas frecuencias. Además, f_{c2} es la frecuencia de corte superior para la cual la señal se atenúa en un factor 0.707 (ec. 9.33), y j es la unidad imaginaria, es decir $j = \sqrt{-1}$.

De acuerdo con la ec. [11.2] el factor de amplificación, A_f , para el circuito realimentado será:

$$A_f = \frac{A_o / [1 + j(f/f_{c2})]}{1 - \beta A_o / [1 + j(f/f_{c2})]} = \frac{A_o}{1 - \beta A_o + j(f/f_{c2})} \quad [11.8]$$

Dividiendo el numerador y denominador de la última expresión por $1 - \beta A_o$, el resultado anterior queda como:

$$A_f = \frac{A_{of}}{1 + j(f/f_{c2,f})} \quad [11.9]$$

donde:

$$A_{of} = \frac{A_o}{1 - \beta A_o} \quad [11.10]$$

y

$$f_{c2,f} = f_{c2}(1 - \beta A_o) \quad [11.11]$$

Si comparamos la expresión [11.7] con la [11.9] podemos concluir que en la región intermedia de frecuencias (es decir, cuando $f \ll f_{c2,f}$) el factor de amplificación del circuito realimentado, se mantiene constante e igual al valor calculado en la ec. [11.2], es decir no está afectado por los condensadores de acoplo u otras capacitancias parásitas. En cambio, en la región de altas frecuencias sufre una atenuación en un factor dado por: $[1+j(f/f_{c2,f})]^{-1}$. Este resultado implica que la frecuencia $f_{c2,f}$ juega el papel de *frecuencia de corte* para la región de altas frecuencias cuando el circuito está realimentando. Esta nueva frecuencia de corte está relacionada con la del circuito sin realimentación, f_{c2} , a través del factor $(1-\beta A_o)$, ec. [11.11]. Así pues, en realimentación negativa, es decir con $\beta < 0$, la frecuencia de corte se hace más elevada. Es más, el aumento de la frecuencia de corte ocurre en la misma proporción que lo que disminuye la ganancia por efecto de la realimentación, de acuerdo con la siguiente relación:

$$A_{of} f_{c2,f} = A_o f_{c2} \quad [11.12]$$

tal como se deduce a partir de las ecuaciones [11.10] y [11.11].

Utilizando un análisis similar para el comportamiento del amplificador realimentado en la región de frecuencias bajas, se obtiene también una nueva frecuencia de corte, $f_{c1,f}$:

$$f_{c1,f} = \frac{f_{c1}}{1 - \beta A_o} \quad [11.13]$$

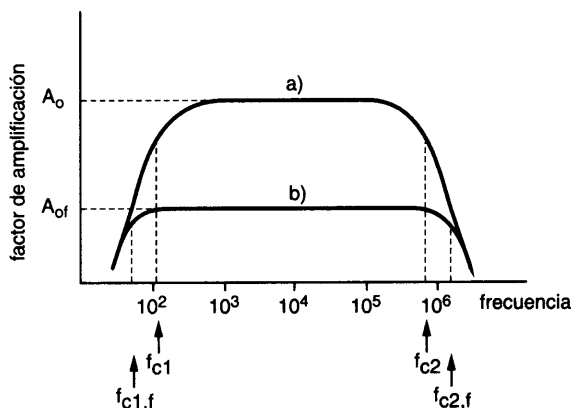


Fig. 11.4. Variación con la frecuencia del factor de amplificación de un circuito sin realimentación (curva a) y con realimentación (curva b).

Esto quiere decir que, en la región de frecuencias bajas, la frecuencia de corte del circuito realimentado también varía en el factor $1/(1-\beta A_o)$. Así pues, en realimentación negativa la frecuencia de corte inferior se desplaza hacia un valor más bajo.

En amplificadores típicos de audiofrecuencia, en los que $f_{c2} \gg f_{c1}$, el valor de la anchura de banda prácticamente coincide con el de f_{c2} , es decir $\Delta f \approx f_{c2}$. Por tanto, la relación [11.12] se puede interpretar en el sentido de que **el producto de la ganancia por la anchura de banda del amplificador es la misma con o sin realimentación**. Esto quiere decir que en realimentación negativa disminuye la ganancia y aumenta la anchura de banda, ambos en el factor $1-\beta A$, mientras que en realimentación positiva ocurre lo contrario.

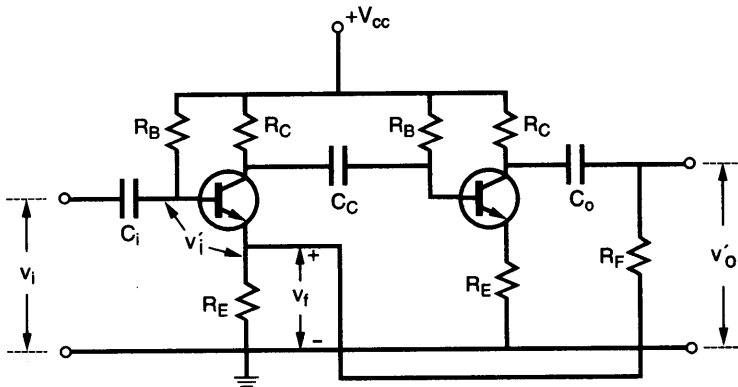
En la fig. 11.4 se comparan la curva de respuesta de un amplificador típico con la correspondiente a la del amplificador realimentado. Según se observa, en esta última curva la ganancia se mantiene constante en un rango más amplio que la del amplificador sin realimentación, y no presenta además las pequeñas oscilaciones que ocasionalmente aparecen como consecuencia de las inestabilidades del circuito no realimentado.

11.3. EJEMPLOS DE CIRCUITOS AMPLIFICADORES REALIMENTADOS

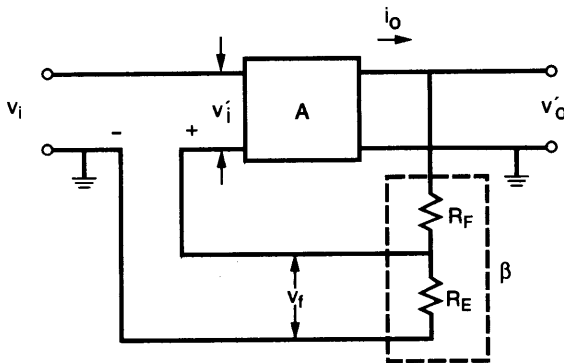
Quizás, algunos de los circuitos estudiados en capítulos anteriores constituyen los mejores ejemplos de amplificadores con realimentación de voltaje. Tal es el caso del **circuito de autopolarización**, desarrollado en la sec. 9.4.1. Según vimos, este circuito lleva una resistencia de emisor común a los circuitos de entrada y salida del amplificador (fig. 9.13). Esto implica una toma de señal paralelo y realimentación en serie, ya que una parte de la señal de salida (la correspondiente a la caída de voltaje en R_E) se superpone a la señal de entrada, pero en oposición de fase. Se produce así un efecto de realimentación negativa que disminuye el factor de amplificación, siendo la reducción tanto mayor cuanto mayor es la resistencia R_E . Sin embargo, el factor de estabilidad del circuito aumenta sensiblemente, según se discutió más arriba (véase también sec. 9.4.1).

El circuito **seguidor de emisor** (apartado 10.2) está basado en los mismos principios, ya que en este caso la resistencia de emisor, R_E , también es común a los circuitos de entrada y salida (fig. 10.4). En este circuito, toda la señal de salida (toma de señal en paralelo) es ahora realimentada en serie en el circuito de entrada, también en oposición de fase. En estas condiciones $\beta = -1$, por lo que el factor de ganancia con realimentación, A_f , se reduce prácticamente a la unidad si el factor A es suficientemente elevado, ya que se cumple entonces: $A_f = A/(1+\beta A) = A/(1+A) \approx 1$. Según vimos, una de las ventajas de esta configuración es el aumento de la resistencia de entrada y la disminución de la resistencia de salida.

En la fig. 11.5a se muestra un circuito amplificador de dos etapas con realimentación, formado por dos amplificadores idénticos en la configuración de emisor común con factor de amplificación, A_v , en cada uno de ellos. El condensador C_c sirve para acoplar la salida del amplificador primero a la entrada del segundo (apartado 10.5). Dado que el amplificador cons-



a)



b)

Fig. 11.5. a) Circuito amplificador de dos etapas, con realimentación negativa.
b) Diagrama de bloques para representar el circuito amplificador anterior.

ta de dos etapas idénticas, el voltaje de salida v_o se encuentra en fase con la señal v_i de la entrada. El amplificador de la fig. 11.5a puede ser representado esquemáticamente por el diagrama de la fig. 11.5b. En esta figura se observa que el voltaje de realimentación $v_f = -\beta v_o$ se toma a través del divisor de tensión formado por las resistencias R_E y R_F , que aparecen

conectadas en serie, y se introduce en oposición de fase con la entrada (realimentación negativa). El voltaje que se amplifica vendrá dado por la diferencia entre v_i y v_f , representado por v'_i en la fig. 11.5a, mientras que el factor de realimentación, β , se puede calcular a través de la relación:

$$\beta = - \frac{v_f}{v'_i} = - \frac{i_o R_E}{i_o (R_E + R_F)} = - \frac{R_E}{R_E + R_F}$$

siendo i_o la corriente que circula en el circuito de salida del amplificador realimentado (sin resistencia de carga). Puesto que en este amplificador el factor de amplificación total en lazo abierto (esto es, sin realimentación) viene dado por $A = A_v^2$ (véase apartado 10.5), nos podemos encontrar fácilmente en la situación en que $A \gg 1$, y por tanto $|\beta| \gg 1/A$. Resulta entonces para el amplificador de la fig. 11.5:

$$A_f = - \frac{1}{\beta} = 1 + \frac{R_F}{R_E}$$

Llegamos así de nuevo a la conclusión de que la ganancia global del amplificador realimentado está determinada únicamente por las resistencias de la red de realimentación, siendo independiente de la ganancia A_v de cada una de las etapas del circuito amplificador.

11.4. AMPLIFICADORES OPERACIONALES

Inicialmente, los amplificadores operacionales fueron denominados así por su utilización en pequeños ordenadores de cálculo analógico, para realizar operaciones básicas de suma o producto, cálculo integral y diferencial, etc. Sin embargo, su versatilidad es tan grande que en la actualidad se utilizan en una gran variedad de equipos electrónicos, tales como fuentes de alimentación, osciladores, convertidores analógico-digitales, etc., siendo comercializados como unidades compactas en forma de circuitos integrados. La razón de su gran utilización se debe en gran parte a que normalmente operan con un circuito de realimentación negativa, y con un factor de amplificación de voltaje en circuito abierto, A , muy elevado. Debido a ello, el factor de realimentación, β , cumple fácilmente la condición $|\beta| \gg 1/A$, por lo que las características del conjunto amplificador dependen muy poco del comportamiento de los transistores y otros componentes que lo forman, según se ha visto anteriormente (véase ec. 11.3). En la actualidad, estos amplificadores se construyen enteramente sobre pastillas de silicio por técnicas de integración de microelectrónica que serán descritas más adelante (cap. XIII). En los apartados que siguen se pretende dar una visión general de los aspectos más esenciales de los amplificadores operacionales, así como de sus aplicaciones en circuitos de cálculo y control analógico.

11.4.1. Características del amplificador operacional

En esencia un amplificador operacional es un amplificador diferencial de ganancia muy elevada. Posee por tanto dos entradas y una salida cuya ganancia es proporcional a la diferencia entre las tensiones aplicadas en cada entrada (véase apartado 10.4). Con objeto de conseguir una ganancia elevada, el amplificador está constituido por varias etapas de amplificación. La etapa diferencial suele estar precedida por amplificadores del tipo JFET, lo cual confiere a cada una de las entradas del amplificador diferencial una alta resistencia de entrada (del orden de decenas o centenas de megaohmios). Asimismo, existe también una etapa de potencia precediendo a la etapa final de salida. Esta etapa final de salida está basada en un amplificador seguidor de emisor (apartado 10.2), por lo que el conjunto adquiere una resistencia de salida muy baja (varias decenas de ohmios). Esto implica que el amplificador operacional es capaz de suministrar una señal de salida con corriente elevada. Las etapas de amplificación son de acoplamiento directo, es decir sin condensadores, lo cual permite que el amplificador sea capaz de operar con señales tanto en continua como en alterna, siempre que la frecuencia no sea muy elevada (del orden de 10-100 kilohercios). La utilización de acoplamiento directo, así como el hecho de poseer una ganancia elevada (alrededor de 100.000), hace que el circuito amplificador sea muy inestable. De ahí surge la necesidad de recurrir a otros circuitos adicionales incluidos en el propio amplificador, cuya misión es asegurar la estabilidad del conjunto y conseguir así un funcionamiento adecuado. En lo que sigue, trataremos el amplificador operacional como un bloque "o caja negra" que presenta unas características determinadas, sin entrar en las particularidades del circuito amplificador.

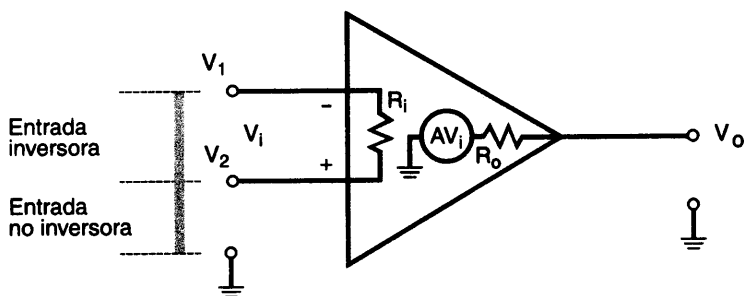


Fig. 11.6. Esquema de un amplificador operacional ideal.

En su forma más común, un amplificador operacional consta de dos *terminales de entrada*, V_1 y V_2 , denominados *inversor* (-) y *no inversor* (+), respectivamente, y un único terminal de salida, V_o . Nótese que la terminología utilizada para las entradas es la misma que la del amplificador diferencial. El circuito se suele representar esquemáticamente por un triángulo, tal como se indica en fig. 11.6. En esta representación se omiten por simplicidad los terminales que conectan el amplificador a la fuente de alimentación de los transistores de cada etapa amplificadora. A veces, no siempre, se añade de forma explícita un terminal de tierra, común a los dos terminales de entrada y al de salida del amplificador. El esquema de la fig. 11.6 incluye también la resistencia de entrada, R_i , entre los dos terminales de entrada, así como el generador voltaje para la señal de salida, AV_i , y la correspondiente resistencia serie, R_o , típicos de cualquier amplificador.

El *amplificador operacional ideal* presenta las siguientes características:

- i) El voltaje de salida del amplificador, V_o , es siempre proporcional a la diferencia entre los voltajes de entrada, V_1 y V_2 , esto es $V_o = A(V_2 - V_1)$. Esta condición implica que si $V_2 = V_1$ entonces $V_o = 0$, incluso a pesar de que A puede ser muy elevado. Según vimos en el apartado 10.4, para conseguir que se cumpla esta condición es preciso que el factor de amplificación del modo común sea cero.
- ii) El factor de amplificación de voltaje A es muy alto, idealmente infinito, e independiente de la frecuencia. Se supone además que el amplificador está libre de las inestabilidades ocasionadas por la elevada ganancia.
- iii) La resistencia de entrada medida entre los dos terminales de entrada, R_i , es muy alta e idealmente se puede considerar infinita.
- iv) La resistencia de salida, R_o , es muy baja y se aproxima a cero.

Como consecuencia de estas características, **en el amplificador operacional ideal no existe consumo de corriente por los terminales de entrada cuando se conecta una señal en el dispositivo** ya que la resistencia de entrada es infinita.

Para entender mejor el papel que juegan los dos terminales de entrada, V_1 y V_2 , supongamos que el potencial aplicado al terminal inversor V_1 es variable mientras que el potencial aplicado al terminal no inversor V_2 permanece constante. Entonces el voltaje de salida V_o cambia con polaridad opuesta a V_1 . Sin embargo, cuando V_2 cambia y V_1 permanece constante, V_o varía en el mismo sentido que V_2 . Una primera aplicación del amplificador operacional es como *dispositivo comparador*, ya que V_o se anula solamente cuando V_1 iguala exactamente a V_2 . Si $V_1 > V_2$ entonces V_o será negativo y si $V_1 < V_2$ entonces V_o será positivo. Los comparadores, en los que el amplificador operacional no opera con circuito de realimentación, se utilizan mucho en aplicaciones en las que se pretende la detección del nivel de una señal, como son los convertidores analógico-digitales.

11.4.2. Realimentación operacional: Concepto de tierra virtual

En los circuitos con amplificadores operacionales la forma más utilizada de realimentación, la denominada *realimentación operacional*, se implementa en la forma indicada en la fig. 11.7, es decir, mediante una resistencia, R_f , en el circuito de realimentación. En esta configuración, la señal que se pretende amplificar, V_i , se introduce en la entrada inversora a través de una resistencia R , mientras que la entrada no inversora se encuentra conectada a

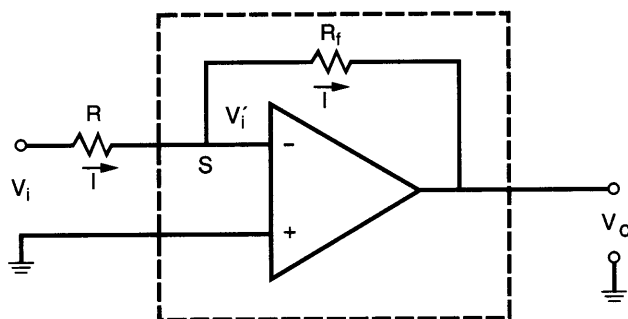


Fig. 11.7. Amplificador operacional con realimentación en la entrada inversora (entrada no inversora unida a tierra).

potencial de tierra ($V \equiv 0$). El amplificador operacional sin realimentación se considera que es ideal, y por tanto con un factor de amplificación de voltaje y una resistencia entre los terminales de entrada infinitos.

Una característica muy importante del circuito de la fig. 11.7 es que el punto S, denominado a veces *punto suma*, constituye un punto de tierra «virtual», ya que su potencial V'_i es prácticamente igual a cero. En efecto, cuando el circuito está en operación, la presencia de una señal finita a la salida, V_o , implica que la diferencia de tensión a la entrada del amplificador, V'_i , debe ser muy próxima a cero, ya que $V_o = AV'_i$, con $A \approx \infty$. Puesto que el terminal no inversor se encuentra conectado a tierra, se tendrá que el potencial de la entrada inversora (-) será igualmente el de tierra.

El término «virtual» se utiliza para indicar que este punto se comporta como si estuviera a potencial de tierra ($V'_i = 0$), aunque de hecho no esté conectado a tierra. Sin

embargo, esto no ocurre para efectos de la corriente ya que la tierra virtual no actúa como sumidero de corriente. Esto es debido a que el amplificador operacional, al tener impedancia infinita, no consume corriente. Conviene señalar además que el terminal inversor actúa como tierra virtual solamente cuando la entrada no inversora está conectada a tierra. La función de la resistencia R que precede al terminal inversor en el circuito de la fig. 11.7 es precisamente para evitar que la señal de entrada pase directamente al terminal inversor del amplificador, ya que en ese caso quedaría cortocircuitada al potencial cero de la tierra virtual.

Si en el circuito de la fig. 11.7 se aplica una señal, V_i , a la entrada existirá una corriente I a través de la resistencia de entrada R . Según se ha mencionado, la impedancia de entrada del amplificador operacional es infinita, y por tanto la corriente I se desvía hacia la salida a través de la resistencia R_f . Así pues, de la igualdad de la corriente en las resistencias R y R_f , se tiene:

$$\frac{V_i - V'_i}{R} = \frac{V'_i - V_o}{R_f}$$

y dado que $V'_i = 0$, el factor de amplificación de voltaje con realimentación operacional será:

$$A_f = \frac{V_o}{V_i} = - \frac{R_f}{R} \quad [11.14]$$

Es decir, el factor de amplificación con realimentación operacional sólo depende del cociente entre las resistencias externas al amplificador operacional R_f y R . El signo negativo en la ecuación [11.14] significa que la señal de salida varía en sentido opuesto con respecto a la de entrada, es decir, V_o es negativo si la señal V_i es positiva, o a la inversa. En el caso de señales alternas, esto significa un desfase de 180° entre la señal de entrada y la de salida. Por esta razón, al circuito de la fig. 11.7 se le conoce también como *inversor de voltaje*.

Desde el punto de vista de un amplificador con realimentación, si comparamos la ecuación [11.14] con la fórmula [11.3], que da el factor de amplificación con realimentación para el caso $\beta \gg 1/A$, resulta para la realimentación operacional que el factor β coincide con el cociente R / R_f .

Es interesante observar que, aunque la resistencia de entrada del amplificador operacional es prácticamente infinita, éste no es el caso en la configuración del amplificador operacional realimentado. En efecto, la resistencia de entrada del circuito de la fig. 11.7 será igual al voltaje de entrada, $V_i - V'_i$, dividido por la corriente de entrada, I , esto es $(V_i - V'_i)/I$, cociente que aproximadamente vale R , ya que $V'_i \approx 0$. Así, por ejemplo, si $R = 10\text{K ohm}$, $R_f = 100\text{K ohm}$ y $V_i = 2\text{V}$, de la ecuación [11.14] se tiene que el voltaje de salida es $V_o = -20\text{V}$, la resistencia de entrada es 10K ohm y la corriente de entrada, igual al voltaje de entrada dividido por la resistencia de entrada, viene dada por: $I = (2\text{V}) / (10\text{K ohm}) = 0.2\text{ mA}$.

11.5. APLICACIONES DE LOS AMPLIFICADORES OPERACIONALES

11.5.1. Circuito amplificador sumador

Una de las ventajas del amplificador inversor de voltaje es la posibilidad de introducir varias entradas a la vez en el punto S. Se obtiene así el *circuito sumador*, representado en la fig. 11.8 para el caso de tres entradas. Dado que el punto S se encuentra a potencial de tierra, las corrientes a través de cada una de las entradas vendrán dadas por:

$$I_1 = V_1 / R_1, \quad I_2 = V_2 / R_2, \quad I_3 = V_3 / R_3$$

Según hemos visto, el punto S constituye una tierra virtual y por tanto no consume corriente (es decir, no se deriva ninguna corriente hacia el interior del amplificador operacional). La corriente, I , a través de la resistencia R_f debe ser por tanto igual a la suma de las tres corrientes, esto es, $I = I_1 + I_2 + I_3$. Esta corriente provoca una caída de potencial en la resistencia R_f dada por:

$$V_i - V_o = -V_o = IR_f$$

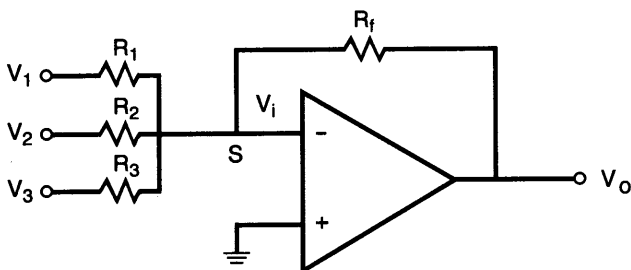


Fig. 11.8. Circuito amplificador sumador.

ya que $V_i \equiv 0$ (punto de tierra virtual). De las ecuaciones anteriores se desprende que la salida V_o es igual a la suma de los voltajes introducidos a la entrada multiplicados cada uno de ellos por un cierto factor, esto es:

$$V_o = - \left(\frac{R_f}{R_1} V_1 + \frac{R_f}{R_2} V_2 + \frac{R_f}{R_3} V_3 \right) \quad [11.15]$$

Debido al resultado de la ec. [11.15], a este *circuito* se le denomina a veces *sumador ponderado* y se puede utilizar, por ejemplo como un mezclador de audio para superponer los sonidos procedentes de varios micrófonos. También se utiliza en circuitos de conversión digital a analógica, así como en circuitos de control donde la señal de salida es igual a la suma de las señales de entrada, cada una con un factor de peso diferente.

Si cada una de las resistencias de entrada es igual a R y si además $R = R_f$, entonces de la ecuación [11.15] se obtiene simplemente:

$$V_o = - (V_1 + V_2 + V_3) \quad [11.16]$$

En el caso particular en que los valores de todas las resistencias a la entrada sean iguales al producto $R_f \cdot n$, donde n es el número de entradas se tiene entonces:

$$V_o = - \frac{V_1 + V_2 + V_3 + \dots + V_n}{n} \quad [11.17]$$

El circuito sirve entonces para calcular el valor medio de varias señales de entrada.

11.5.2. Circuito amplificador restador

El circuito amplificador restador se muestra en la fig. 11.9. Según se observa, en este circuito ninguna de las dos entradas al amplificador operacional se encuentra conectada directamente a tierra. Denominemos V_i al voltaje tanto de la entrada inversora como de la no inversora, ya que la diferencia entre los voltajes de cada una de las dos entradas es prácticamente cero. Además, puesto que no fluye corriente a través de los terminales de entrada al amplificador se tiene:

$$\frac{V_1 - V_i}{R_1} = \frac{V_i - V_o}{R_2} \quad [11.18]$$

y

$$\frac{V_2 - V_i}{R_1} = \frac{V_i}{R_2} \quad [11.19]$$

siendo V_1 y V_2 las señales introducidas en las entradas del circuito (véase fig. 11.9). Restando de la ecuación [11.19] la [11.18] resulta para el voltaje de salida:

$$V_o = \frac{R_2}{R_1} (V_2 - V_1)$$

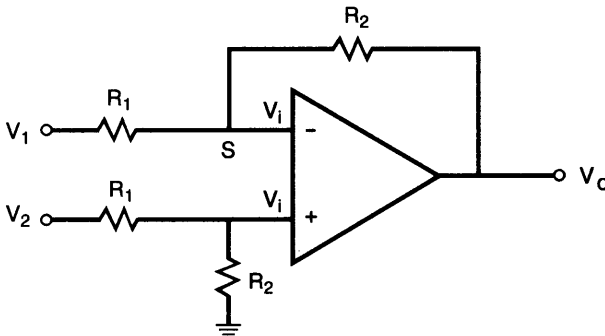


Fig. 11.9. Circuito amplificador restador.

Si el circuito se diseña de forma que $R_2 = R_1$ se tiene entonces que $V_o = V_2 - V_1$, es decir, el voltaje de salida es igual a la diferencia entre los dos voltajes de entrada.

11.5.3. Circuito integrador

El circuito de la fig. 11.10 se puede utilizar para integrar una señal de voltaje $V_i(t)$ variable con el tiempo aplicada a la entrada. Este circuito está provisto de una resistencia R en el terminal inversor, que se utiliza para introducir la señal de entrada, manteniendo el terminal no inversor unido a tierra. El lazo de realimentación está formado por un condensador, C , de capacidad conocida. Efectivamente, el voltaje, V , desarrollado al cabo de un cierto tiempo

entre las placas del condensador es igual a la carga, $Q(t)$, que se acumula en las placas dividido por la capacidad, C , del condensador. Como la carga instantánea es igual a la integral de la corriente con respecto al tiempo se tiene que:

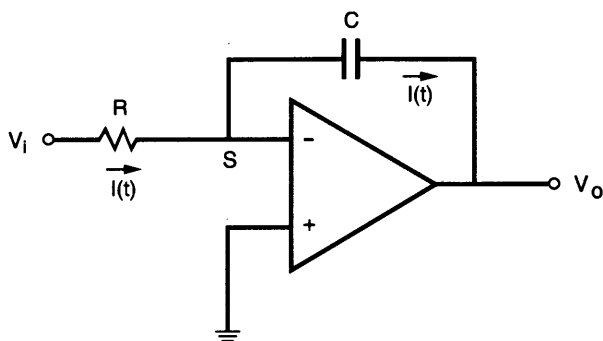


Fig. 11.10. *Circuito amplificador integrador.*

$$V = \frac{1}{C} \int I(t) dt \quad [11.20]$$

En esta configuración, la corriente instantánea que pasa por el condensador es la misma que la que atraviesa la resistencia R de la fig. 11.10, por lo que la corriente $I(t)$ viene dada por:

$$I(t) = \frac{V_i(t)}{R} \quad [11.21]$$

Dado que el punto S está a potencial de tierra, el voltaje V a través del condensador es $-V_o$ (para el sentido indicado de la corriente). Sustituyendo este valor en la ecuación [11.20] y teniendo en cuenta [11.21] se tiene:

$$V_o = - \frac{1}{RC} \int V_i(t) dt \quad [11.22]$$

Aunque el voltaje de salida del integrador representa el área bajo la curva V_i en función de t , los integradores se utilizan principalmente como generadores de rampas y en filtros electrónicos. Así, se puede generar una rampa triangular a partir de una señal de voltaje de entrada constante, o desfazar 90° una señal sinusoidal. El dieléctrico del condensador utilizado debe de ser de muy alta calidad, ya que cualquier corriente de fugas que lo atraviere da un valor erróneo de la integral. Asimismo, el condensador debe contener un interruptor en paralelo para poder hacer cero la carga en sus placas siempre que sea necesario.

11.5.4. Circuito diferenciador

Se puede obtener la derivada de una función representada por un voltaje, $V_i(t)$, variable mediante el circuito de la fig. 11.11, el cual utiliza un condensador, C , conectado por un extremo a la entrada del terminal inversor, con el terminal no inversor unido a tierra. La señal de entrada se introduce a través del otro extremo del condensador. Comparando este circuito con el de la fig. 11.10 para el integrador, se observa que se han intercambiado la resistencia y el condensador. Conforme V_i varía con el tiempo se produce una corriente $I(t)$, también variable, a través del condensador C y de la resistencia de realimentación R_f , ya que como hemos visto, el punto S se encuentra siempre a potencial de tierra, y además no se deriva corriente hacia el interior del amplificador. Si denominamos $Q(t)$ a la carga sobre las placas del condensador, la corriente a través de éste viene dada por:

$$I(t) = \frac{dQ}{dt} = C \frac{dV_i(t)}{dt}$$

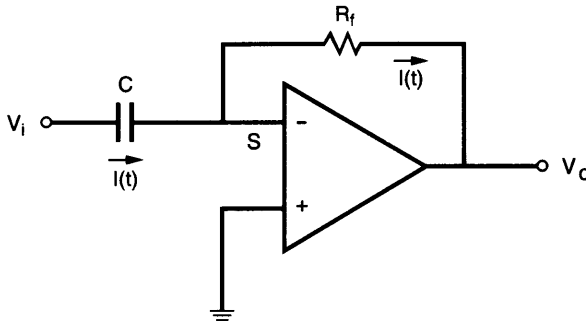


Fig. 11.11 Circuito amplificador diferenciador.

Por tanto:

$$V_o = - R_f I(t) = - R_f C \frac{dV_i}{dt} \quad [11.23]$$

La ecuación [11.23] indica que la señal de salida del circuito de la fig. 11.11 es proporcional a la derivada de la señal de entrada. Es de esperar que este circuito sea muy sensible al ruido eléctrico en forma de pulsos casi instantáneos de voltaje, por lo que en su utilización es necesario el apantallamiento eléctrico del circuito.

11.5.5. Circuito medidor de corriente

Como es bien sabido, un medidor de corriente se introduce en serie en la línea en que se pretende hacer la medida, por lo que el medidor es tanto mejor cuanto menor es su resistencia de entrada. Por otra parte, según vimos en el apartado 11.4, los amplificadores operacionales, en su configuración de realimentación operacional, presentan una resistencia de entrada cuyo valor coincide con el de la resistencia conectada en el terminal inversor de entrada. Esta resistencia puede ser elegida con un valor tan bajo como se quiera, incluso cero en el caso ideal. El medidor de corriente de la fig. 11.12 se basa en estos principios. En este circuito la corriente a medir, I , entra directamente por el terminal inversor hacia el punto S, que constituye una tierra virtual. Esta corriente se deriva a través de la resistencia situada en el lazo de

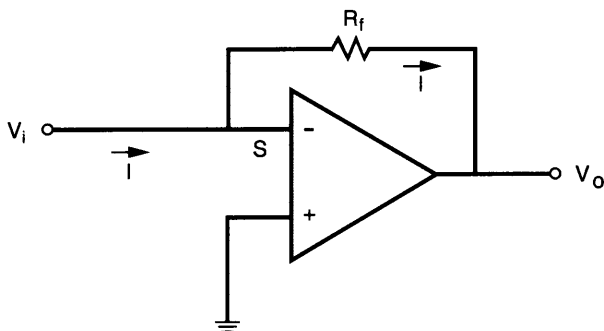


Fig.11.12. Circuito medidor de corriente.

realimentación, por lo que la caída de tensión a través de ella, V_o , vendrá dada por:

$$V_o = - IR_f \quad [11.24]$$

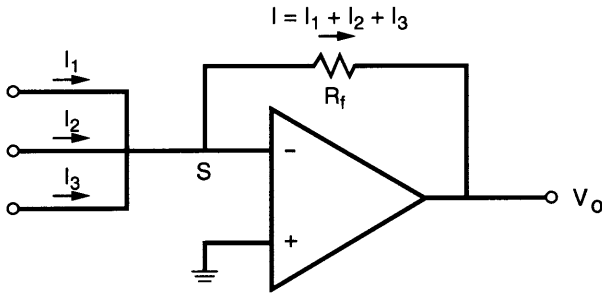


Fig.11.13. Circuito de medida de la suma de tres corrientes.

Por tanto, la medida del voltaje de salida del amplificador operacional da el valor de la corriente de entrada multiplicado por la resistencia de realimentación. A este circuito también se le denomina conversor de corriente a voltaje. El método descrito es mucho más práctico para medir corrientes que el de la medida directa de la caída de voltaje a través de una resistencia de valor conocido insertada en el circuito. Así, utilizando un valor alto de R_f en el circuito de la fig. 11.12 se pueden medir corrientes tan pequeñas como 10^{-12} A con bastante exactitud. Una de las características de este circuito es que la resistencia de entrada es prácticamente nula, aunque obliga a que el punto donde se hace la medida de la corriente esté a potencial cero. Puesto que el punto S de la fig. 11.12 sirve también como punto de suma para varias corrientes (I_1 , I_2 e I_3 , por ejemplo), el circuito se puede utilizar para obtener un voltaje de salida, V_o , proporcional a la suma de las corrientes I_1 , I_2 e I_3 presentes en el terminal de entrada, es decir:

$$V_o = - (I_1 + I_2 + I_3)R_f \quad [11.25]$$

La fig. 11.13 muestra el circuito que permite realizar la operación suma de corriente.

11.5.6. Circuito seguidor de voltaje

A menudo es necesario aislar un generador de una señal de voltaje de un determinado aparato de medida que posea una resistencia de entrada baja, ya que puede causar una carga de corriente excesiva al generador. Para evitar este efecto pernicioso se puede emplear un circuito basado en el amplificador operacional, indicado en fig. 11.14. En este circuito se introduce

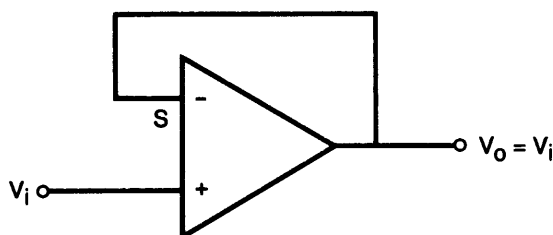


Fig. 11.14. Circuito seguidor de voltaje.

la señal por el terminal de entrada no inversor del operacional y todo el voltaje de salida es realimentado al terminal inversor (punto S). Este punto no constituye ahora una tierra virtual y por tanto su potencial puede ser diferente de cero. De hecho, el voltaje en el punto S coincide con el voltaje en el terminal de salida, V_o , y al mismo tiempo con el voltaje en el terminal de entrada, V_i . Es obvio pues que $V_o = V_i$, es decir, el voltaje de salida "sigue" al voltaje de entrada, y el factor de amplificación es igual a la unidad. Todas estas características son comunes al amplificador seguidor de emisor descrito en el apartado 10.2. La resistencia de entrada de este circuito se corresponde con la del terminal no inversor, y puede ser de varias decenas de megaohmios. Por otra parte, la resistencia de salida es muy baja, alrededor de unos pocos ohmios, ya que coincide con la del amplificador operacional.

El comportamiento del seguidor de voltaje hace que este circuito sea muy apropiado en la medida de señales de voltaje que estén acompañadas de una alta resistencia, es decir, en aquellos casos en los que el generador del circuito equivalente tenga una resistencia serie elevada. En estas circunstancias, si se utiliza un medidor de voltaje de tipo convencional, puede ocurrir que la resistencia de entrada del medidor sea mucho mas baja que la del genera-

dor de la señal. La señal a medir se reduce entonces enormemente debido a la caída de tensión en la resistencia interna del propio generador.

En la fig. 11.15 se muestra un ejemplo de medida de la carga acumulada en las placas de un condensador, C , mediante un circuito seguidor de voltaje. Supongamos que el condensador se carga hasta alcanzar un cierto voltaje V_i . Si queremos conocer el valor del voltaje entre las placas del condensador conectando un voltímetro de baja calidad, y por tanto de resistencia de entrada pequeña, el condensador se puede descargar a través del voltímetro (nótese que el condensador puede ser considerado en este ejemplo como una fuente de señal con resistencia serie de valor muy elevado). Sin embargo, se puede evitar la descarga conectando el condensador a una etapa seguidora de voltaje, con una resistencia de entrada superior a la resistencia equivalente del condensador. El consumo de corriente en el terminal de entrada del operacional es entonces muy bajo, dando el amplificador una salida, V_o , igual a la del potencial, V_i , que se pretende medir. De este modo, es posible utilizar un voltímetro sencillo a la salida del operacional para medir V_o . El voltímetro queda así aislado de la fuente que proporciona la señal de entrada, sin producir ningún efecto perturbador sobre ésta.

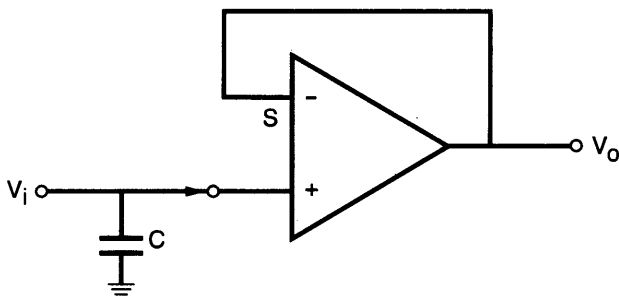


Fig. 11.15. Seguidor de voltaje empleado para medir la diferencia de voltaje entre las placas de un condensador, C .

11.5.7. Circuito amplificador logarítmico (*)

En la fig. 11.16 se muestra un circuito amplificador con realimentación operacional, en el cual se ha conectado en el lazo de realimentación un transistor bipolar con la base conectada a tierra. Como veremos, este tipo de amplificador se utiliza para obtener una señal de salida

que sea proporcional al logaritmo de la señal de entrada. Interesa realizar esta operación cuando la señal de entrada tiene un espectro amplio de variación, por ejemplo, de varias décadas, ya que entonces el rango de variación de la salida queda reducido dentro de un margen mucho más estrecho.

Obsérvese en la fig. 11.16 que el colector del transistor está conectado a la entrada inversora del amplificador operacional (punto S de tierra virtual) mientras que la base está conectada directamente a tierra. En estas condiciones se puede decir que el colector y la base están prácticamente corto-circuitadas. De acuerdo con la ec. [6.16], la corriente del colector I_C en un transistor se puede aproximar por $I_C \approx \alpha_{dc} I_E$, siendo α_{dc} la ganancia en corriente del transistor, con un valor próximo a la unidad. A su vez, la unión de emisor se comporta como un diodo polarizado en directo a la tensión V_o (siempre que V sea negativa), por lo que la corriente de emisor está relacionada con el voltaje V_o de salida mediante la ecuación del diodo, ec. [3.36]:

$$I_E = I_o [\exp (qV_o / kT) - 1]$$

Así pues:

$$I_C \approx \alpha_{dc} I_o \exp (qV_o / kT) \quad [11.26]$$

donde se ha despreciado el término unidad en comparación con el término exponencial. La corriente I_C a su vez es igual a la que circula a través del terminal de entrada, cuyo valor viene

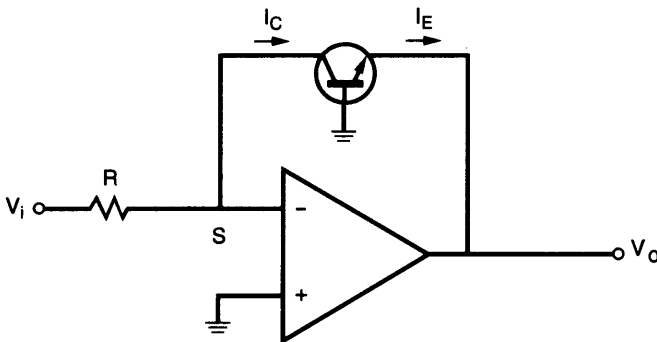


Fig. 11.16. Amplificador logarítmico.

dado por V_i/R , siendo V_i el voltaje introducido en el terminal de entrada. Tendremos por tanto:

$$V_i \approx R I_o \exp (qV_o / kT) \quad [11.27]$$

en la cual se ha supuesto que $\alpha_{dc} \approx 1$. Tomando el logaritmo en ambos miembros de la ec. [11.27] se tiene:

$$V_o = \frac{kT}{q} \ln \frac{V_i}{R I_o} = \frac{kT}{q} \ln \frac{V_i}{R} - \ln I_o \quad [11.28]$$

De la ecuación [11.28] se aprecia que la señal de salida es proporcional al logaritmo de la señal de entrada, tal como se había anticipado más arriba.

11.5.8. El amplificador instrumental (*)

En esta sección trataremos un tipo de amplificador diferencial comúnmente utilizado para aceptar gran variedad de señales en cada uno de los terminales de entrada suministrando una salida única. Por esta razón se le denomina *amplificador instrumental*. El circuito correspondiente se muestra en la fig. 11.17, en el cual se observa a la entrada un amplificador diferencial formado por los operacionales A_1 y A_2 . Los voltajes de entrada V_1 y V_2 se introducen en los terminales no inversores de estos operacionales. Esto implica que las entradas inversoras de los operacionales A_1 y A_2 estarán también a los voltajes V_1 y V_2 , respectivamente, por lo que la corriente a través de R_1 será:

$$I = (V_2 - V_1) / R_1 \quad [11.29]$$

La corriente I indicada en la fig. 11.17 que atraviesa la resistencia R_1 y la que existe en las dos resistencias R_2 del circuito es la misma, ya que las impedancias de entrada de los operacionales son infinitas. Así pues, la diferencia de voltajes $V'_2 - V'_1$ a la salida de los operacionales A_1 y A_2 será igual a la suma de las caídas de voltaje a través de las dos resistencias R_2 y de la resistencia R_1 de la figura, es decir:

$$V'_2 - V'_1 = I (R_1 + 2R_2) \quad [11.30]$$

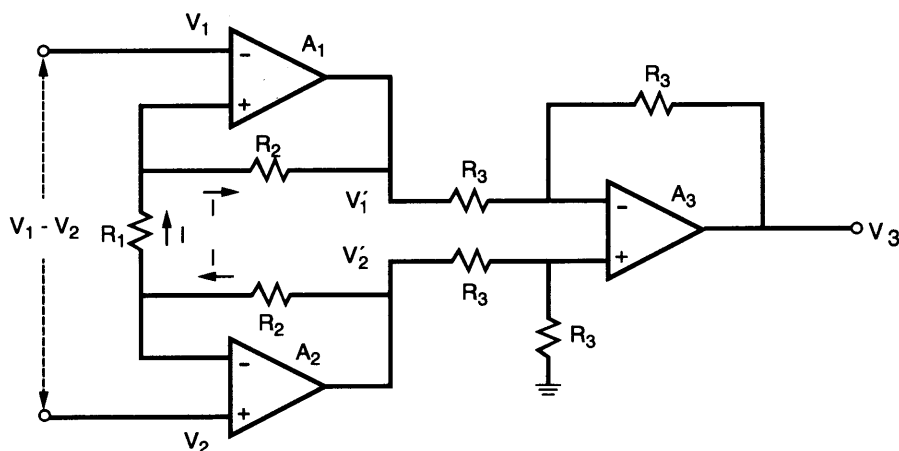


Fig. 11.17. Amplificador diferencial instrumental.

Por tanto, de las ecuaciones [11.29] y [11.30] se tiene:

$$V'_2 - V'_1 = (V_2 - V_1) \left(1 + \frac{2R_2}{R_1}\right) \quad [11.31]$$

Esta tensión es la que se aplica ahora a la entrada del operacional A_3 en su configuración de amplificador sustractor (apartado 11.5.2). En consecuencia la salida V_3 vendrá dada por:

$$V_3 = V'_2 - V'_1 = (V_2 - V_1) \left(1 + \frac{2R_2}{R_1}\right) \quad [11.32]$$

El amplificador instrumental descrito presenta una serie de características que le hacen muy útil, entre ellas la de tener una entrada diferencial con una salida única, una gran impedancia de entrada y la posibilidad de escoger el factor de amplificación variando solamente el valor de la resistencia R_1 . Por último, el factor de amplificación de señales del modo común (es decir, la parte de la señal que aparece en las dos entradas, según vimos en el apartado 10.4) es muy bajo, de forma que, por ejemplo, el ruido a 50 Hz producido por la red, que aparece como dos señales en fase en los terminales de entrada, se amplifica mucho menos que otras componentes no comunes de las señales de entrada.

11.5.9. Cálculo analógico mediante amplificadores operacionales

Las computadoras analógicas basadas en los amplificadores operacionales realimentados, se pueden utilizar para la resolución de muchos problemas físicos tales como aquellos que exigen la solución de una ecuación diferencial. De este modo, se pueden simular electrónicamente una gran variedad de problemas físicos.

Como ejemplo, tomemos la ecuación diferencial del oscilador armónico amortiguado que describe las vibraciones de una masa m unida a un muelle de constante de fuerza k y que se mueve en un medio viscoso de constante de amortiguamiento b . Se supone además que sobre la partícula actúa una fuerza $F(t)$ dependiente del tiempo. La ecuación diferencial que describe la posición de x viene dada por:

$$\frac{d^2x}{dt^2} + \frac{b}{m} \frac{dx}{dt} + \frac{k}{m} x = \frac{F(t)}{m} \quad [11.33]$$

Supongamos que disponemos de una señal de voltaje proporcional a d^2x/dt^2 y veamos cómo, con la computadora analógica representada por el circuito de la fig. 11.18, se puede resolver la ec. [11.33]. La señal d^2x/dt^2 es primeramente integrada por el operacional integrador 1 de la figura en el cual se elige R y C_1 de forma que $RC_1 = 1$ s. Según vimos en el apartado 11.5.3 la señal obtenida a la salida es $-dx/dt$, la cual puede ser nuevamente integrada, obteniéndose la variable x a la salida del operacional integrador 2. Simultáneamente, la tensión del terminal 1 se introduce en el sumador 4 (para el cual se hace $R/R_1 = b/m$) junto con la tensión proporcional a $F(t)/m$ proveniente de un terminal de entrada. Se obtiene así a la salida del operacional 4 una señal igual a $(b/m)(dx/dt) - F(t)/m$. A su vez, esta salida es introducida junto con la variable x del terminal 2 en el sumador 3 que tiene una resistencia R_2 a la entrada, de valor $R/R_2 = k/m$. A la salida del operacional 3 se obtiene la suma:

$$- \frac{k}{m} x - \frac{b}{m} \frac{dx}{dt} + \frac{F(t)}{m}$$

cuyo valor ha de ser igual a d^2x/dt^2 , y que se vuelve a introducir a la entrada del operacional 1. El circuito por tanto resuelve la ecuación diferencial [11.33], obteniéndose la solución en el terminal 2 de la fig. 11.18 siempre que se introduzcan los valores apropiados de las condiciones iniciales ($t = 0$) de las variables x y dx/dt . Las condiciones iniciales se pueden introducir aplicando en los condensadores C_1 y C_2 un voltaje proporcional a $(dx/dt)_0$ y x_0 , tal como se indica en la fig. 11.18. Para obtener la solución con las condiciones iniciales apropiadas se abren los interruptores S_1 y S_2 al mismo tiempo que se cierra S_3 . La variación del desplazamiento x (en el terminal 2) y de la velocidad dx/dt (en el terminal 1) pueden observarse conec-

tando un osciloscopio en los terminales correspondientes del circuito de la fig. 11.18. Haciendo las resistencias R_1 , R_2 y R variables se puede hacer una simulación electrónica para estudiar cómo influyen los parámetros del oscilador e incluso las condiciones iniciales en la solución final del oscilador armónico.

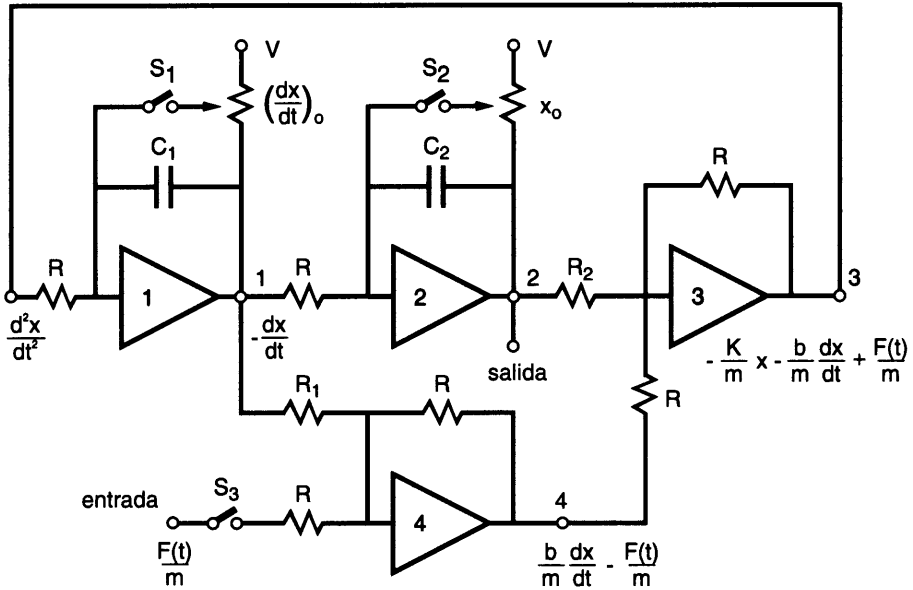
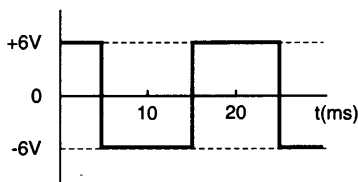


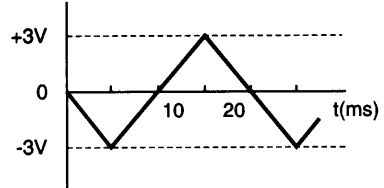
Fig. 11.18. Esquema de una computadora analógica, empleada para resolver la ecuación diferencial de un oscilador armónico.

CUESTIONES Y PROBLEMAS

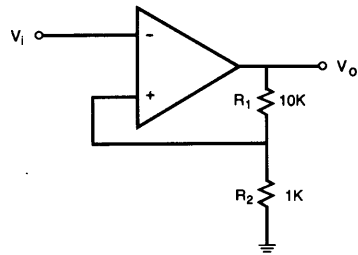
- 11.1** Supongamos un amplificador cuyo factor de amplificación sin realimentación tiene una inestabilidad de hasta un 40%. A partir de él se quiere construir un amplificador realimentado con una estabilidad en la amplificación mejor que el 1% y un factor de amplificación de 30. Calcular el factor de amplificación en lazo abierto y el factor de realimentación necesarios.
- 11.2** Un amplificador, cuya ganancia en circuito abierto es de 40, tiene un lazo de realimentación negativa del 15%. Si la ganancia en circuito abierto se incrementa en un 10%, ¿cuánto varía la ganancia del circuito con realimentación?
- 11.3** Para el amplificador de dos etapas con realimentación negativa de la fig. 11.5, hallar aproximadamente el factor de amplificación y el factor de realimentación (suponer que $R_E = 3K \text{ ohm}$, $R_F = 40K \text{ ohm}$ y que los transistores de silicio presentan unas características típicas de este material).
- 11.4** En el circuito con amplificador operacional de la fig. 11.7, se tiene: $R = 1K \text{ ohm}$, $R_f = 100K \text{ ohm}$. Determinar el factor de amplificación y el valor de V_o para $V_i = 0,1 \text{ V}$ (valor de pico) en los siguientes casos: i) El amplificador operacional es ideal. ii) El amplificador operacional tiene un factor de amplificación $A = 10^6$. iii) Idem si $A = 10^5$.
- 11.5** El integrador de la fig. 11.10 tiene $C = 1 \mu\text{F}$ y $R = 10 K \text{ ohm}$. Calcular el voltaje de salida para la señal de entrada de la figura, con amplitud 6 V y período 20 ms.



- 11.6** El circuito diferenciador de la fig. 11.11 tiene $C = 1 \mu\text{F}$ y $R = 10\text{K ohm}$. Calcular el voltaje de salida para la señal de entrada de la figura, con una amplitud 3V y período 20 ms.



- 11.7** Para el amplificador no inversor de la figura, calcular el factor de amplificación y la señal de salida correspondiente a una señal de entrada de 1V. Comparar el circuito de la figura con el del seguidor de voltaje.



CAPITULO XIII

TECNOLOGIA DE DISPOSITIVOS MICROELECTRONICOS

La reducción en tamaño de los dispositivos electrónicos ha llevado a la denominada *tecnología de microelectrónica*, que constituye el diseño y la fabricación de circuitos formados por muchos componentes (principalmente transistores) de tamaño extremadamente pequeños. Muy a menudo estos *circuitos* se denominan *integrados*, ya que el conjunto de componentes electrónicos está montado sobre una placa de silicio o chip (cuya dimensión lateral típica es de unos pocos milímetros), que no puede ser dividida en partes independientes. De ahí que a veces a estos circuitos se les denomine también con el nombre genérico de *monolíticos* (del latín, cuyo significado literal es "una piedra"). A principios de los años 60 se fabricaban los circuitos con una integración a pequeña escala (SSI) con menos de unos 100 componentes (entre transistores, diodos, condensadores y resistencias) por chip, para pasar a últimos de esa década a la integración a gran escala (LSI) con un número de componentes entre 1.000 y 10.000 por chip. A mediados de los 70 comenzó la denominada integración a escala muy grande (VLSI) con más de 10.000 componentes por chip para llegar en la actualidad a la integración de varios millones de componentes por chip (ULSI o integración a escala ultra grande). La preparación de estos circuitos requiere el uso de una gran variedad de técnicas que van desde la obtención de la oblea de silicio en forma de monocristal hasta la deposición de capas metálicas para formar los contactos y pistas de interconexión. En este capítulo trataremos de dar una visión general de las técnicas básicas de fabricación de estos dispositivos microelectrónicos.

13.1. CIRCUITOS INTEGRADOS MONOLITICOS

A partir del descubrimiento del transistor en 1947, que reemplazó a las tradicionales válvulas de vacío, uno de los objetivos de los fabricantes ha sido producir circuitos electrónicos cada vez más pequeños, menos costosos, con menor consumo de energía y con respuesta más rápida. Al principio los transistores se fabricaban individualmente, pero alrededor de 1958 se desarrolló la idea de fabricar el conjunto de dispositivos (transistores, diodos, condensadores, etc.) sobre una pieza única de silicio, también llamada "chip", ocupando aproximadamente el mismo espacio que el de un transistor discreto.

Así, a partir de cada oblea de silicio, con un diámetro típico de entre 10 y 20 cm, se pueden fabricar cientos de circuitos integrados, cada uno conteniendo miles y en algunos casos hasta varios millones de componentes. En principio, se podría pensar que la fabricación de circuitos complejos con muchos componentes interconectados en un único chip de pequeño tamaño podría implicar serios riesgos económicos y técnicos. Sin embargo, de hecho los circuitos microelectrónicos son más fiables y relativamente más económicos. Esto se debe a

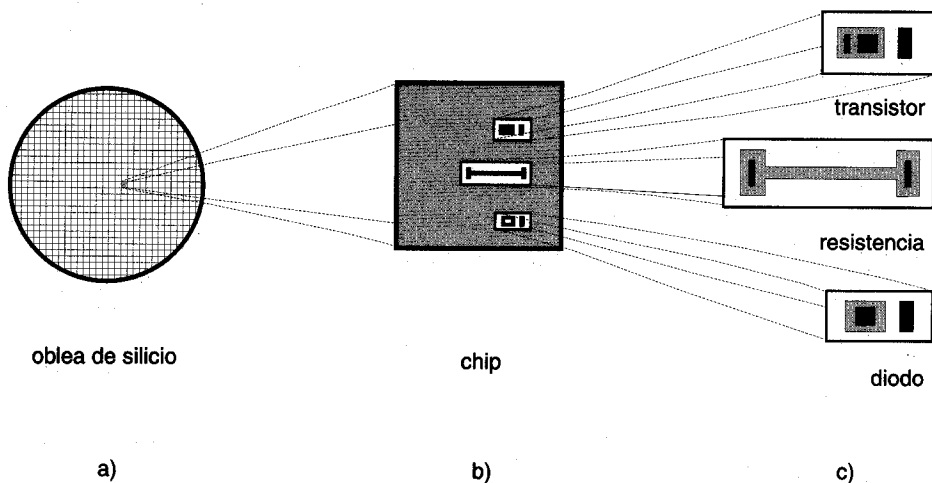


Fig. 13.1. La figura muestra una oblea de silicio (a) a partir de la cual se pueden obtener cientos de chips idénticos (b). A su vez cada chip está formado principalmente por multitud de transistores y diodos, pero también pueden haber elementos pasivos como resistencias (c).

que se puede fabricar simultáneamente sobre una misma oblea de silicio una cantidad elevada de circuitos idénticos. La mayor fiabilidad se debe a su vez a que los componentes y sus interconexiones se efectúan sobre un sustrato rígido (sin partes móviles) y único.

Otras ventajas de los circuitos integrados se derivan de su pequeño tamaño y peso que los hace aptos para utilizarse en aplicaciones tales como vehículos espaciales o potentes ordenadores, aunque cada vez se emplean más en electrónica de consumo (televisores, teléfonos, automóviles, etc.). Uno de los aspectos más interesantes de la miniaturización está relacionado con la velocidad de propagación de señales eléctricas en los circuitos, ya que la transmisión de señales está limitada por la velocidad de la luz (unos pocos decímetros por nanosegundo). Por tanto, si se pretende construir un ordenador con una gran velocidad de respuesta su tamaño deberá ser pequeño. Además, las capacidades e inductancias parásitas que aumentan el tiempo de transferencia de señales son muy pequeñas en los circuitos integrados. La fig. 13.1 muestra un esquema de una oblea de silicio formada por cientos de chips en cada uno de los cuales existe un cierto número de transistores, diodos y otros componentes, ocupando todos ellos un pequeño espacio.

Las tecnologías para la fabricación de circuitos integrados son una extensión de la *tecnología planar* utilizada en la producción de transistores discretos, según se verá en el próximo apartado. En esencia esta tecnología consiste en la difusión de impurezas en áreas previamente seleccionadas de una oblea de silicio con el fin de crear zonas de carácter p ó n, así como uniones p-n. Todos los componentes se colocan sobre la misma superficie de una de las caras de la oblea, y por ello a esta tecnología se la denomina planar.

13.2. TECNOLOGIA PLANAR

La mayoría de los transistores que se fabrican actualmente en unidades discretas se preparan a partir de obleas de silicio mediante un proceso de difusión de impurezas sobre la superficie de la oblea. Describiremos la tecnología planar para el caso de un transistor de silicio tipo npn, situado sobre la superficie de la oblea, ocupando un área pequeña. Según se indica en la fig. 13.2, el proceso de fabricación se inicia con una oblea ó pieza de silicio monocristalino de tipo n, sobre cuya superficie se forma previamente una capa de óxido por calentamiento en atmósfera oxidante (a). Como es sabido, el óxido de silicio es relativamente impermeable al paso de impurezas, de ahí que la misión de esta capa sea la de proteger la superficie de la oblea en determinadas zonas para evitar la difusión de impurezas durante el proceso de dopaje del silicio que se realiza posteriormente. Mediante la técnica de fotolitografía, que se verá después, se elimina el óxido por ataque químico en áreas prefijadas formando una ventana en cada una de ellas (b). Estas áreas delimitan el espacio que ha de ocupar el transistor. Efectivamente, sobre la superficie del silicio que ha quedado al descubierto se realiza el proceso de dopaje mediante difusión de impurezas de tipo p para formar la base del transistor

(c). Se forma después una nueva capa de óxido sobre el silicio por calentamiento en una atmósfera oxidante y se abren nuevas ventanas en el óxido formado, de menor tamaño que las anteriores (d), con objeto de realizar un nuevo dopaje y difundir impurezas de tipo n de la zona de emisor (e). Finalmente se oxida de nuevo la superficie y se abren nuevas ventanas aplicando una vez más las técnicas fotolitográficas, para efectuar sobre el espacio abierto de las ventanas los contactos metálicos al colector, base y emisor (f).

Mediante este proceso es posible fabricar a partir de una sola oblea y en un proceso único cientos o miles de transistores en pequeñas placas. Esto se hace simplemente abriendo las correspondientes ventanas en diferentes áreas del semiconductor y ejecutando simultáneamente en cada ventana la secuencia de operaciones mencionadas. Cada una de las áreas que contiene el transistor se separa posteriormente cortando la oblea en pequeñas placas. Finalmente, cada placa individual se monta en un soporte y se sueldan conexiones a los contactos de emisor, base y colector constituyendo cada placa un transistor.

Refiriéndonos a los circuitos integrados, los procesos necesarios de fabricación son muy similares a los ya descritos para la tecnología planar. La fig. 13.3a muestra un ejemplo de un circuito simple formado por una resistencia, un diodo, un transistor y un condensador. La estructura de las distintas capas delgadas que forman este circuito cuando se fabrica sobre una

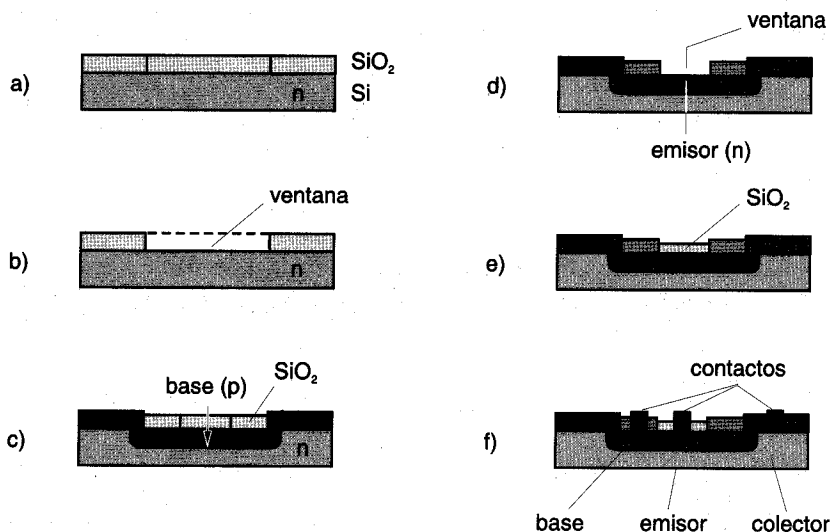


Fig. 13.2. Proceso de fabricación de un transistor planar npn de silicio. Cada oblea de silicio da como resultado cientos o miles de transistores individuales que son fabricados simultáneamente.

pieza única de silicio viene dada en la fig. 13.3b. Básicamente, para la fabricación de un circuito integrado de esta naturaleza, se pueden distinguir las siguientes etapas de preparación:

- i) En primer lugar está el crecimiento de la barra o lingote de silicio perfectamente cristalizado. Estas barras suelen tener hoy día un diámetro entre 10 y 20 cm y una longitud de un metro, aproximadamente. La barra se corta en obleas circulares de alrededor de un cuarto de milímetro de espesor, y a continuación son pulidas cuidadosamente para evitar imperfecciones superficiales (sec. 13.3).
- ii) Las obleas así obtenidas se utilizan solamente como sustrato o soporte de los circuitos integrados, ya que normalmente se deposita sobre la oblea una fina capa de silicio de unas pocas micras de espesor y con la misma estructura cristalina que el sustrato (capa epitaxial). Esta capa juega un papel fundamental en la preparación de circuitos integrados ya que es en ella donde realmente se producen los distintos componentes activos y pasivos. Más adelante veremos cómo se hace la preparación de esta capa epitaxial (sec. 13.4).

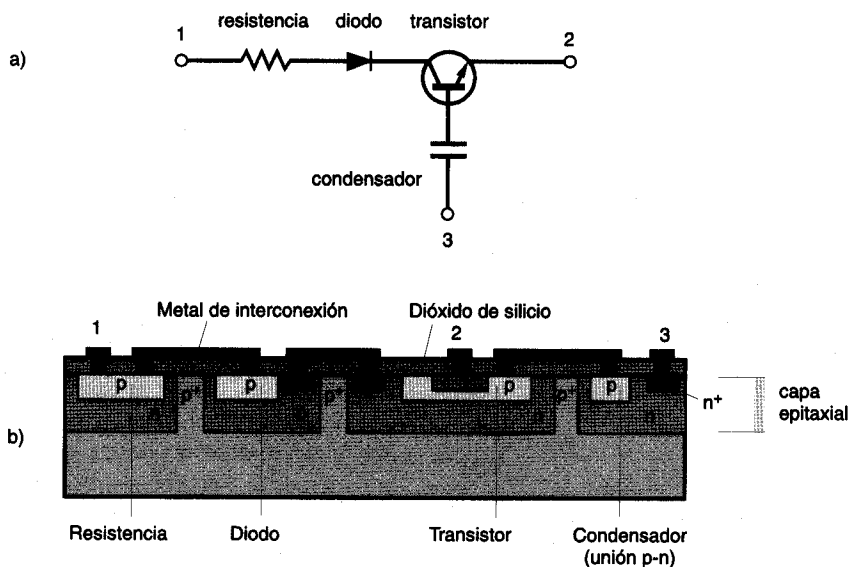


Fig. 13.3. a) Ejemplo de un circuito electrónico sencillo para ser fabricado según la tecnología planar. b) Estructura de capas del circuito anterior, fabricado según la tecnología planar, mostrando las diferentes capas: epitaxial, óxido de silicio y la capa metálica para formar los contactos al semiconductor y las pistas conductoras.

- iii) Durante la producción de los circuitos integrados es preciso oxidar la superficie del silicio y hacer después un ataque selectivo en diferentes zonas o áreas, prefijadas de antemano. Se forman así ventanas en el óxido que dejan al descubierto pequeñas áreas de la superficie del silicio. Estos procesos de oxidación y ataque se alternan frecuentemente con otros procesos de formación de las diferentes capas dopadas que componen el circuito (sec. 13.5). El objetivo final es efectuar el dopaje en las zonas donde el óxido ha sido atacado y el silicio ha quedado al descubierto.
- iv) La eliminación selectiva por agentes químicos del óxido de silicio en áreas o zonas prefijadas se efectúa por un proceso denominado grabado. Previamente es preciso delimitar las zonas del óxido que es preciso atacar o eliminar, y esto se realiza mediante otro proceso adicional conocido como fotolitografía. Este último proceso se basa fundamentalmente en la utilización de máscaras para recubrir y proteger, mediante un material orgánico tipo resina, aquellas áreas del óxido en las que no es preciso efectuar el ataque (sec. 13.6).
- v) El proceso más importante en la fabricación de circuitos integrados es probablemente el de dopaje mediante la difusión de impurezas hacia el interior de la capa epitaxial (sec. 13.7). Este proceso se realiza bien sea en hornos apropiados, a temperaturas de unos 1000 °C, o bien mediante un método diferente conocido como implantación. En el primer caso, las fuentes de impurezas pueden ser gases, líquidos o sólidos que se ponen en contacto con las zonas de silicio en las que se ha eliminado el óxido protector. En el método de implantación, un haz de iones formado por los iones de fósforo, arsénico, boro, etc., es enviado con energías entre 100 y 200 KeV contra la superficie del silicio, introduciéndose en él a profundidades que dependen esencialmente de la energía de los iones incidentes.
- vi) Un aspecto importante en la producción de circuitos integrados es el del aislamiento eléctrico entre componentes electrónicos situados en posiciones contiguas. Un método de conseguir este aislamiento consiste en abrir ventanas en las zonas de óxido comprendidas entre los componentes para formar regiones con dopaje de tipo opuesto. En el ejemplo de la fig. 13.3b, con una capa epitaxial de tipo n, estas regiones son de tipo p muy dopadas (p^+). De este modo, cada componente se encuentra localizado en una isla de material de tipo n rodeada por una región de tipo p. El aislamiento eléctrico se consigue conectando el sustrato al potencial más negativo del circuito por lo que cada componente queda rodeado por una unión p-n polarizada en inverso.
- vii) Finalmente es necesario depositar contactos metálicos a los semiconductores, así como interconectar los distintos componentes entre sí, utilizando para ello materiales de alta conductividad eléctrica, como el aluminio, cobre, etc. (indicados en negro en la fig. 13.3b). La deposición de contactos de tipo metálico se realiza generalmente mediante procesos de evaporación en vacío del metal correspondiente, o también median-

te otros procesos más complejos en los que el metal se deposita tras un bombardeo con iones positivos muy energéticos que arranca partículas de un cátodo metálico (pulverización catódica) (sec. 13.8).

13.3. CRECIMIENTO DEL SILICIO MONOCRISTALINO

Un requerimiento esencial para la utilización del silicio en la preparación de los circuitos integrados es que el material de partida sea de gran pureza y se encuentre en estado monocristalino. El material monocristalino se obtiene a partir de silicio policristalino de gran pureza, también denominado silicio de grado electrónico. Este silicio a su vez se consigue a partir de sílice (SiO_2) relativamente pura, por reducción química mediante calentamiento con

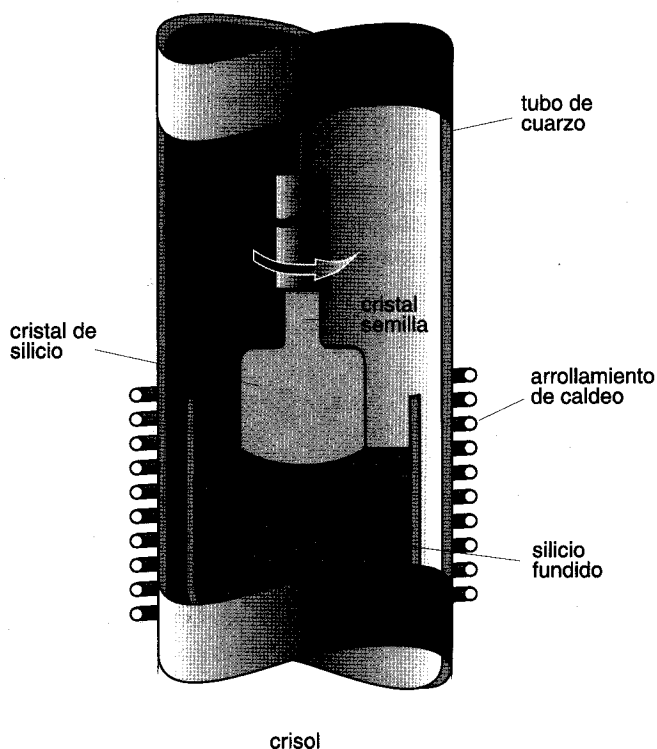


Fig. 13.4. Esquema del sistema experimental utilizado para crecer silicio monocristalino por la técnica de Czochralski.

carbono. El material resultante debe ser purificado de nuevo, para lo cual se trata con cloruro de hidrógeno (HCl). Se forma así triclorosilano (SiHCl_3), que es líquido a la temperatura ambiente, el cual se purifica aún más mediante sucesivas destilaciones. Finalmente el SiHCl_3 se reduce con hidrógeno para formar silicio de grado electrónico en estado policristalino.

El silicio monocristalino utilizado en los circuitos integrados se obtiene a partir del silicio policristalino, el cual es convertido al estado de monocristal mediante la *técnica de Czochralski*, desarrollada en 1917 para la preparación de metales en forma de monocristal. Según esta técnica (fig. 13.4), un pequeño monocristal de silicio de aproximadamente 1 cm de diámetro - denominado *semilla o germen* - se introduce parcialmente en un baño de silicio fundido, el cual se encuentra a una temperatura superior a la de fusión (1415°C). Cuando la parte inferior del cristal se empieza a fundir dentro del silicio líquido, se tira lentamente de la semilla hacia arriba. De este modo, el silicio fundido adherido al monocristal se solidifica tomando la estructura cristalina del cristal semilla. Generalmente se hace rotar el cristal semilla alrededor de su eje conforme es elevado para evitar que las diferencias de temperatura puedan causar una solidificación no homogénea. Simultáneamente, el silicio líquido se remueve para que la concentración de las impurezas añadidas en el dopaje se uniformice. Mediante un cuidadoso control de la temperatura de la solución y del desplazamiento del cristal semilla se puede determinar el diámetro del monocristal resultante.

En el crecimiento del monocristal de silicio mediante la técnica de Czochralski, se añade una cantidad determinada de dopante al silicio en solución. La concentración de impurezas incorporada al monocristal difiere generalmente de la concentración de la solución en la interfase. La razón de estas dos concentraciones se denomina *coeficiente de segregación o distribución*, K , y está dado por:

$$K = C_s / C_l \quad [13.1]$$

donde C_s y C_l son respectivamente las concentraciones de equilibrio del dopante en el sólido y en el líquido. El valor de K depende de la impureza utilizada en el dopaje. En el caso del silicio K vale 0.35 ó 0.80 para las impurezas de fósforo o boro, respectivamente.

La composición del crisol de la fig. 13.4, que contiene el silicio fundido, es importante ya que debe de ser de un material que reaccione lo menos posible con el silicio e introduzca cantidades inapreciables de impurezas no intencionadas. El material más utilizado es el cuarzo (SiO_2), el cual reacciona parcialmente con el silicio fundido, contaminando el silicio monocristalino con concentraciones de unos 10^{17} átomos de oxígeno por cm^3 . Por este motivo, en la actualidad se está considerando la utilización del nitruro de silicio para formar el crisol, ya que este material da lugar a una menor contaminación. Otra impureza que aparece a menudo en el silicio crecido por la técnica de Czochralski es el carbono en concentraciones de unos 10^{16} átomos por cm^3 y suele provenir de los elementos de grafito que contiene el horno.

Un aspecto importante a tener en cuenta es la atmósfera dentro del horno, que suele ser de un gas inerte, tal como el argón, para evitar la oxidación del silicio monocristalino crecido.

13.4. CRECIMIENTO DE LA CAPA EPITAXIAL DE SILICIO

La palabra *epitaxia* proviene del griego y significa "capa ordenada". Efectivamente, en el crecimiento epitaxial se forma una fina capa cristalina de silicio (de 5 a 10 μm de espesor) sobre el sustrato monocristalino. La capa de silicio depositada suele tener un dopaje distinto que el del sustrato. De este modo, el crecimiento epitaxial permite preparar capas con diferentes tipos de impurezas, sin necesidad de recurrir al dopaje por difusión a través de capas ya impurificadas, lo cual requeriría niveles de dopaje muy elevados.

En el proceso de crecimiento epitaxial se hace pasar un compuesto gaseoso de silicio (tetracloruro de silicio, p. e.) sobre una oblea de silicio caliente. Las altas temperaturas hacen que el compuesto de silicio se descomponga bien sea por reacción química o bien por pirólisis. Los átomos de silicio formados llegan al sustrato y se trasladan sobre su superficie hasta alcanzar puntos de nucleación correspondientes a las posiciones que los átomos ocupan en la red cristalina. Por todo ello es muy importante que la superficie del monocristal que actúa de sustrato sea de gran pureza y sin ningún tipo de imperfecciones, tales como dislocaciones. Generalmente, para obtener una rápida velocidad de nucleación la superficie cristalina suele presentar una dirección cristalográfica a unos pocos grados de direcciones principales tales como las $\langle 111 \rangle$ ó $\langle 100 \rangle$. De este modo se presentan escalones en la superficie que actúan como puntos de nucleación.

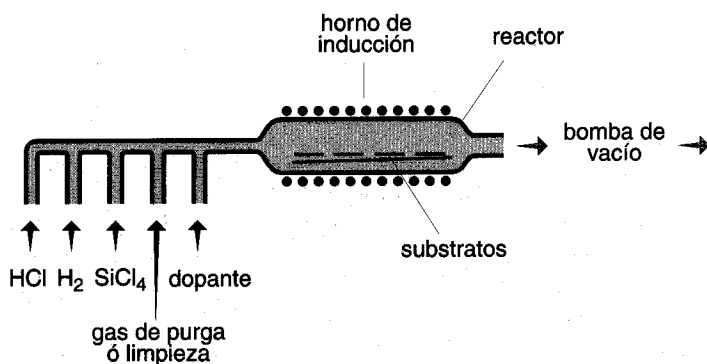
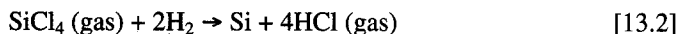
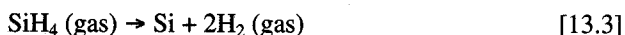


Fig. 13.5. Esquema de un sistema experimental típico para la deposición de capas epitaxiales de silicio sobre sustratos de silicio.

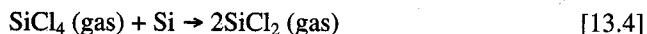
Las reacciones más usuales para la deposición epitaxial del silicio son la reducción por hidrógeno del tetracloruro de silicio a unos 1200 °C:



o bien, la descomposición pirolítica del silano a unos 1000 °C:



La fig. 13.5 muestra un reactor típico utilizado en el crecimiento epitaxial del silicio. La reacción ocurre en el tubo de cuarzo en cuyo interior se colocan los sustratos sobre un susceptor de grafito calentado. Este susceptor de grafito se calienta a la temperatura deseada mediante un sistema de inducción eléctrica de radiofrecuencia. Es interesante observar que la reacción [13.2] es reversible, por lo que si se introduce ácido clorhídrico en forma gaseosa en el sistema la superficie del silicio queda atacada. Este hecho se utiliza para llevar a cabo el proceso de limpieza del silicio previamente a la deposición. El ataque de la superficie del monocristal tiene lugar a través de la reacción:



la cual compite con la reacción [13.2] en el proceso de deposición del silicio.

En la fig. 13.6 se muestra la velocidad de crecimiento de la capa epitaxial de silicio en función de la fracción molar de SiCl_4 en relación a la cantidad total del gas portador, que suele ser hidrógeno. Se puede observar que al principio la velocidad de crecimiento aumenta con la concentración de SiCl_4 hasta alcanzar un máximo. A partir del máximo la velocidad de crecimiento comienza a disminuir debido al ataque del silicio según la reacción [13.4]. El silicio epitaxial se prepara en la región de bajas concentraciones de SiCl_4 donde hay crecimiento positivo, como se indica en la fig. 13.6.

El dopaje de las capas epitaxiales se consigue introduciendo simultáneamente con el compuesto gaseoso de silicio y el gas portador otros gases conteniendo las impurezas adecuadas, tales como la fosfina (PH_3) o el diborano (B_2H_6). El dopaje conseguido de este modo es aproximadamente una función lineal de la razón de las moléculas del gas dopante en relación al gas que contiene los átomos de silicio.

La *epitaxia por haces moleculares* (MBE) es otro proceso que, aún no estando demasiado implantado todavía a escala industrial, posee un gran futuro para la obtención de capas epitaxiales sobre un sustrato monocristalino. Las capas que se crecen por MBE se forman a partir de materiales calentados a altas temperaturas en unas celdas o crisoles fabricados gene-

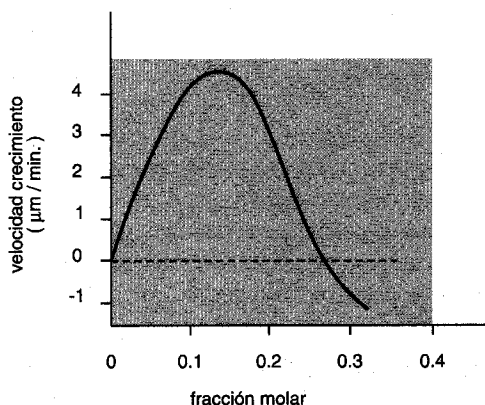


Fig. 13.6. Influencia de la concentración de SiCl_4 en la velocidad de crecimiento de la capa epitaxial de silicio.

ralmente con nitruro de boro. La temperatura de cada horno se ajusta dentro de rangos muy estrechos para obtener la velocidad de evaporación adecuada. Los átomos de material evaporado se depositan sobre la superficie del sustrato, el cual se encuentra en condiciones de ultra alto vacío (10^{-10} Torr de presión) para evitar contaminaciones. En el presente, la técnica de MBE se utiliza más en la deposición de capas epitaxiales de elementos III-V (arseniuro de galio principalmente) que en la tecnología del silicio. La técnica ha sido aplicada, por ejemplo, en la obtención de estructuras periódicas de capas semiconductoras de GaAs con distintos dopantes y un espesor equivalente a unas pocas capas atómicas. Las estructuras así obtenidas se denominan *superredes* y se utilizan tanto en investigaciones de carácter fundamental como en aplicaciones a dispositivos electrónicos y optoelectrónicos.

13.5. FORMACION DE CAPAS FINAS AISLANTES

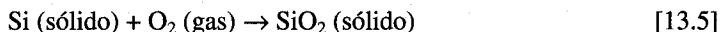
13.5.1. Oxido de silicio

De todos los dieléctricos, el óxido de silicio es el más empleado en la industria de los semiconductores. De hecho, el lugar preponderante que ocupa el silicio entre los semiconductores se debe en gran parte a las buenas propiedades del óxido de silicio, tanto desde el punto de vista eléctrico - es uno de los mejores aislantes conocidos - como en sus

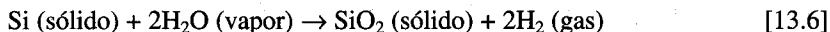
aplicaciones para formar capas tipo barrera para frenar la difusión de dopantes e impurezas a través de ellas. Además, el óxido de silicio se obtiene fácilmente por oxidación térmica de la superficie del silicio. Por el contrario, en el caso del arseniuro de galio la oxidación térmica produce una mezcla de óxidos de galio y arsénico no estequiométricos, de baja calidad como aislantes eléctricos y como protectores de la superficie.

El óxido de silicio que encuentra más aplicaciones en microelectrónica es el dióxido, de fórmula SiO_2 , y constante dieléctrica relativa, $\epsilon = 4$, aproximadamente. Por otra parte, el campo eléctrico de ruptura del SiO_2 es bastante elevado, cerca de 10^7 Vcm^{-1} , lo que le hace idóneo como aislante en la integración de circuitos a muy gran escala donde, debido al pequeño espesor de los componentes, las capas aislantes de los circuitos integrados deben soportar campos eléctricos muy elevados.

Las películas de óxido de silicio utilizadas en estos casos se preparan por oxidación térmica del sustrato de silicio a temperaturas en el rango de 900-1200 °C, bien sea en atmósfera de oxígeno seco o bien en vapor de agua. Las correspondientes reacciones químicas son:



y



En la fig 13.7 se ha representado la variación del espesor de óxido con el tiempo de oxidación a diferentes temperaturas para cada una de las dos reacciones anteriores. Según se observa, la reacción de oxidación del silicio mediante vapor de agua procede con una velocidad mucho mayor (alrededor de un orden de magnitud) que con oxígeno seco.

En la oxidación térmica, las obleas de silicio se colocan verticalmente sobre un soporte de cuarzo (o grafito), en el cual previamente se han practicado varias ranuras de sujeción. Este soporte, a su vez, está situado dentro de un tubo de cuarzo calentado mediante un horno exterior. Por un extremo del tubo de cuarzo se introducen los gases reactantes (oxígeno seco o vapor de agua) y por el otro se cargan las obleas. La operación de carga de muestras debe realizarse bajo condiciones extremas de limpieza, es decir en habitaciones libres de cualquier contaminación de polvo (sala limpia, apartado 13.6).

La reacción para la formación del óxido ocurre en la interfase Si-SiO₂, por lo que las especies oxidantes deben atravesar mediante un proceso de difusión la capa de óxido previamente formada. Utilizando las ecuaciones de Fick (véase sec. 13.7.1) para este proceso de difusión se obtiene la siguiente expresión para la dependencia del espesor de óxido crecido,

d_{ox} , en función del tiempo, t , de reacción:

$$d_{ox} = \frac{A}{2} \left[\left(1 + \frac{t + \tau}{A^2 / 4B} \right)^{1/2} - 1 \right] \quad [13.7]$$

donde A y B son parámetros que dependen del coeficiente de difusión de las especies oxidantes, de su concentración en la superficie y en el interior del óxido y de la velocidad de reacción para la oxidación. El parámetro τ tiene en cuenta los efectos de la capa de óxido inicial presente sobre el silicio (*óxido nativo*).

Para tiempos de oxidación suficientemente cortos, de forma que se cumpla la condición $(t + \tau) \ll A^2 / 4B$, se tiene:

$$d_{ox} = \frac{B}{A} (t + \tau) \quad [13.8]$$

resultando una dependencia lineal del espesor del óxido con el tiempo. En este caso, la reacción en la interfase Si-SiO₂ es la que limita la velocidad de crecimiento, por lo que ésta depende de la naturaleza de los enlaces superficiales, y por tanto, de la orientación cristalográfica de la superficie del silicio. Así, por ejemplo, la cara $\langle 100 \rangle$ se oxida más lentamente que la cara $\langle 111 \rangle$. Para tiempos largos, en los que $t \gg A^2 / 4B$, resulta de la ec. [13.7]:

$$d_{ox} = [B (t + \tau)]^{1/2} \quad [13.9]$$

En este caso la dependencia del espesor con el tiempo es de tipo parabólico. Además, la reacción de oxidación está limitada por el proceso de difusión de las especies oxidantes a través de la capa de óxido previamente formada. Esto explica que no exista dependencia de la velocidad de oxidación con la orientación de la superficie del silicio, según se ha observado experimentalmente.

En la oxidación térmica del silicio es importante considerar que, como consecuencia del crecimiento del óxido a partir del silicio, la interfase Si-SiO₂ se desplaza hacia el interior del silicio. Denominando d_{ox} el espesor del óxido crecido y d_1 el espesor por debajo de la superficie original del silicio (fig.13.8), a partir de los datos de la densidad del silicio y de su óxido, se calcula fácilmente que $d_1 = 0.46 d_{ox}$.

Un aspecto importante relacionado con el crecimiento es que la capa de óxido de silicio obtenida en oxígeno seco presenta mejores propiedades eléctricas que la preparada a partir del vapor de agua, aunque, como ya se mencionó, en el primer caso es necesario un tiempo alrededor de un orden de magnitud mayor para obtener un espesor dado. Por ello los óxidos muy finos, como los utilizados en la puerta de los transistores MOS ($100 \text{ \AA} \leq d_{ox} \leq 1000 \text{ \AA}$), se

crecen en oxígeno seco, mientras que aquellos más gruesos empleados para el aislamiento eléctrico ($d \approx 5000 \text{ \AA}$) se obtienen en atmósfera de vapor de agua.

Las fórmulas desarrolladas anteriormente para el crecimiento del espesor del óxido de silicio en función del tiempo están muy de acuerdo con los resultados experimentales para óxi-

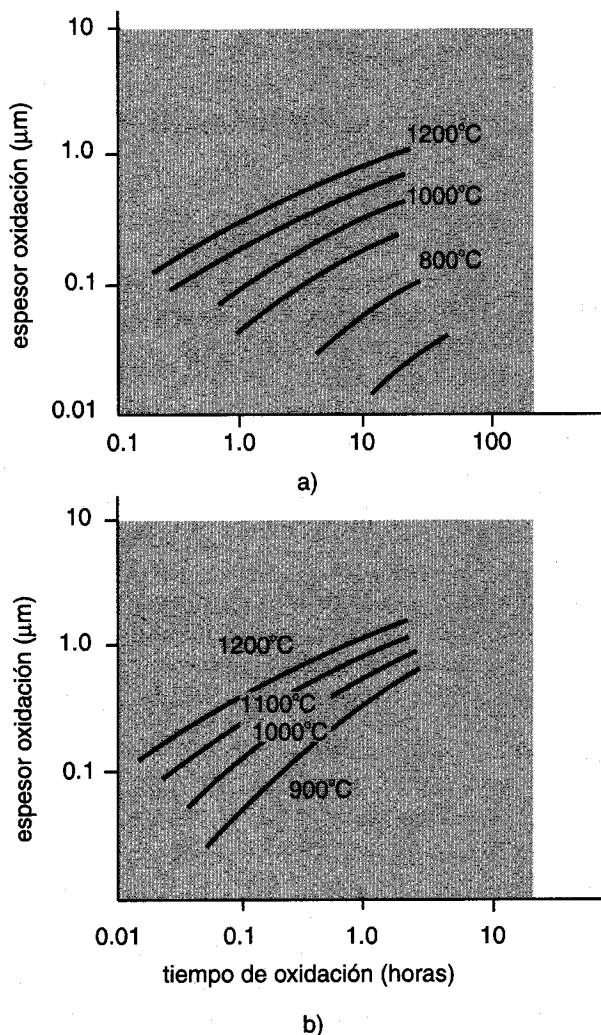


Fig. 13.7. Crecimiento del óxido de silicio a varias temperaturas en atmósfera de oxígeno seco (a) y en atmósfera de vapor de agua (b).

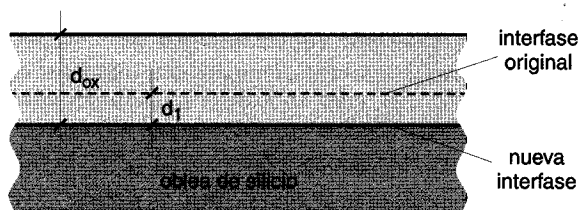


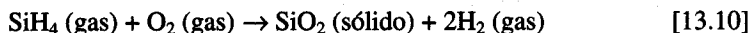
Fig. 13.8. Efecto del crecimiento del óxido en la superficie del silicio.

dos relativamente gruesos ($d > 500 \text{ \AA}$). Sin embargo, en la tecnología VLSI y más aún en la submicrónica es a veces necesario la utilización de óxidos tan finos como unos 100 \AA . En estos óxidos ultrafinos se ha observado que el crecimiento con el tiempo se puede aproximar relativamente bien mediante una ley parabólica. Diversas teorías han sido desarrolladas para explicar este fenómeno atribuido a las elevadas tensiones de compresión a que es sometido el óxido durante los primeros estadios de crecimiento, lo que a su vez produce como consecuencia una fuerte reducción del coeficiente de difusión para las especies oxidantes.

La calidad dieléctrica de las capas finas de óxido de silicio se determina por dos parámetros. Uno de ellos es el campo eléctrico de ruptura del óxido, que debe aproximarse a los 10^7 Vcm^{-1} . Este campo se determina aplicando una diferencia de potencial creciente entre dos electrodos, uno de ellos depositado sobre la superficie del óxido y el otro sobre la cara opuesta del sustrato de silicio. Se observa entonces el voltaje a partir del cual la corriente aumenta por encima de un cierto nivel. El otro parámetro viene determinado por la concentración de contaminantes presentes en el óxido, especialmente iones móviles de sodio, que originan importantes inestabilidades en el funcionamiento de los transistores tipo MOS. La concentración de contaminantes formados por iones móviles se obtiene mediante la medida de la variación de la capacidad con el voltaje (curvas C-V) en la estructura de tipo MOS formada al depositar un contacto metálico sobre el óxido (apartado 7.3). Las medidas se efectúan a temperaturas relativamente elevadas, manteniendo el terminal de puerta a una polaridad determinada. Después se cambia la polaridad de los terminales y se mide de nuevo la curva C-V. A partir del desplazamiento en el eje de voltajes entre las dos curvas es posible la determinación de la concentración de cargas móviles contaminantes.

A pesar de sus ventajas, la oxidación térmica no puede emplearse cuando se utiliza el óxido de silicio como aislante entre pistas conductoras a distintos niveles, o en la tecnología del arseniuro de galio. Otro inconveniente de la oxidación térmica está relacionado con el hecho de que en los dispositivos microelectrónicos con resolución por debajo de la micra, los procesos a altas temperaturas tienden a redistribuir las concentraciones de dopantes difundidos previamente, originando desplazamientos en la localización de las uniones p-n y cambios en los dopajes del material masivo. Todo ello hace muy deseable la puesta a punto de procesos de deposición directa del óxido de silicio a temperaturas más bajas que las de la oxidación térmica.

Capas finas de óxido de silicio obtenidas a temperaturas bajas se pueden formar a través de una técnica de deposición basada en la reacción de gases en fase vapor. La técnica se conoce como *deposición química en fase vapor o CVD*¹. Las deposiciones a baja temperatura, alrededor de 400 ó 500 °C, se realizan mediante la reacción del silano y del oxígeno, de acuerdo con la ecuación:



Esta reacción se puede llevar a cabo en un reactor a la presión atmosférica, pero si se quiere obtener óxidos con una mejor calidad se utiliza el CVD a *baja presión o LPCVD*, siendo entonces la velocidad de deposición de unos 100 Å min⁻¹. Para temperaturas de deposición algo más altas, hasta unos 700 °C, en lugar del silano se emplea el tetraetilortosilicato (TEOS) en forma líquida. Este compuesto se evapora primero, antes de entrar en el horno, y luego se descompone por efecto del calor dando óxido de silicio y una serie de gases que son evacuados del sistema. En estos sistemas de deposición las obleas de silicio se sitúan en el interior de un reactor, calentado por un horno, a través del cual pasan los gases reaccionantes. El producto de la reacción, el SiO₂ en este caso, aparece en forma de depósito sobre la superficie de las obleas.

13.5.2. Nitruro de silicio

El nitruro de silicio, Si₃N₄, que presenta una resistividad a la temperatura ambiente alrededor de 10¹² ohm cm, un campo de ruptura de 10⁷ Vcm⁻¹ y una constante dieléctrica relativa cerca de 8 (aproximadamente el doble de la del óxido de silicio), está recibiendo últimamente mucha atención para aplicaciones en microelectrónica. Una propiedad interesante del nitruro de silicio es la gran resistencia que opone a la difusión de impurezas (mayor que la del SiO₂), y en particular a los principales contaminantes de los dispositivos

¹ Nota: Los acrónimos CVD y LPCVD proceden de los términos en inglés: "Chemical Vapor Deposition" y "Low Pressure Chemical Vapor Deposition".

semiconductores, como son los iones de sodio. Debido a ello el Si_3N_4 es un material muy apropiado para su utilización como última capa, con fines pasivantes o de protección, en dispositivos de estado sólido y también como barrera de difusión en los procesos fotolitográficos de grabado necesarios para la producción de circuitos integrados (apartado 13.6). Otra aplicación importante del nitruro de silicio es la de aislante alternativo al SiO_2 en las puertas de los transistores MOS ya que, al tener una constante dieléctrica doble a la del óxido, la tensión umbral de funcionamiento del transistor se reduce a la mitad.

Sin embargo, la interfase $\text{Si-Si}_3\text{N}_4$ no es de tanta calidad como la interfase Si-SiO_2 , que es la que presenta menor concentración de estados superficiales y de densidad de carga atrapada (véase apartado 7.4). Por ello, a menudo se emplean en los transistores MOS estructuras del tipo $\text{Si-SiO}_2\text{-Si}_3\text{N}_4\text{-metal}$, con un espesor del óxido de unos 200 Å y del nitruro de 800 Å. La estructura anterior también es la base de las memorias de lectura alterable eléctricamente (EAROM) utilizadas en microelectrónica, en las cuales el espesor del óxido es de tan sólo unos 30 Å y el del nitruro unas diez veces superior. En estas unidades de memoria los electrones atraviesan el óxido por efecto túnel y se quedan atrapados en la interfase $\text{SiO}_2\text{-Si}_3\text{N}_4$, constituyendo así una unidad de memoria.

Las capas finas de nitruro de silicio se pueden depositar por LPCVD a unos 750 °C a partir de la reacción del silano SiH_4 (o bien diclorosilano, SiH_2Cl_2) y del amoníaco NH_3 , utilizando nitrógeno o hidrógeno como gas portador. Las películas así crecidas son bastante estequiométricas y densas (alrededor de 3.0 g cm^{-3}). A menudo, el proceso de CVD se lleva a cabo en el interior de un reactor en el que se provoca una descarga eléctrica (plasma) que acelera la reacción. En este proceso, denominado *CVD asistido por plasma* o PECVD², el nitruro se deposita a tan sólo unos 300 °C por lo que esta técnica generalmente se utiliza para obtener capas de Si_3N_4 empleadas en la etapa final de la fabricación, esto es, como encapsulante de componentes electrónicos. Un inconveniente del nitruro de silicio depositado por PECVD es que presenta una elevada concentración de átomos de hidrógeno, hasta el 20% del total de átomos, lo cual reduce sus características aislantes. Estos átomos de hidrógeno proceden de los gases reaccionantes, como el silano y el amoníaco, los cuales sufren una descomposición parcial durante la deposición, dando lugar a radicales -SiH y -NH que quedan incorporados en el nitruro en forma de subproductos.

13.6. LITOGRAFIA

Muchos de los procesos seguidos en la fabricación de un circuito integrado requieren el ataque químico del óxido que recubre al silicio en ciertas zonas, bien sea para hacer la difusión del material dopante en el semiconductor, o bien para depositar las capas metálicas utilizadas como contactos. Para conseguir este objetivo es necesario efectuar previamente un

² Nota: Acrónimo del inglés: "Plasma Enhanced Chemical Vapor Deposition".

proceso fotolitográfico para delimitar las áreas en las que es preciso realizar el ataque. Básicamente, la técnica consiste en la transferencia de imágenes de formas geométricas delineadas sobre una máscara (opaca a la radiación utilizada) a una fina capa de un material sensible a esa radiación, denominada *resina*³, que cubre la superficie del semiconductor. El patrón para una determinada máscara se dibuja primero a un tamaño normal y se reduce luego por procesos fotográficos en un factor alrededor de 1000. Así por ejemplo, si dos líneas conductoras deben estar separadas por una distancia de 5 μm en el circuito integrado, en la máscara estarán separadas por 5 mm.

Hoy día, gran parte de los equipos litográficos para la producción de circuitos integrados utilizan la radiación ultravioleta, con una longitud de onda de alrededor de 0.3 μm , aunque como veremos luego, para circuitos de muy alta integración puede ser necesario otro tipo de radiación. En la fig. 13.9 se ilustra el proceso de litografía óptica. La oblea con la capa de SiO_2 que se pretende atacar se cubre con una resina sintética fotosensible a la radiación - también denominada *fotorresina* (fig. 13.9a). Una vez que el patrón de la máscara ha sido transferido a una placa fotográfica de vidrio se ilumina la capa de resina con radiación ultravioleta (UV) a través de la placa. De este modo, la radiación incide solamente sobre determinadas zonas de la fotorresina (fig. 13.9b). Por efecto de la radiación UV, la fotorresina se polimeriza debajo de aquellas zonas de la placa transparentes a la radiación, haciéndose insoluble en ciertas soluciones orgánicas (el tricloroetileno, por ejemplo). Por tanto, la inmersión posterior en uno de estos agentes permite disolver aquellas zonas de la fotorresina no expuestas a la radiación UV (fig. 13.9c). La fotorresina no disuelta es después calentada a unos 150 $^{\circ}\text{C}$ para endurecerla y hacerla resistente al ataque de los ácidos que se aplican seguidamente para eliminar el óxido en las zonas que han quedado expuestas. La oblea se introduce entonces en una solución basada en ácido fluorhídrico que ataca solamente el óxido de silicio no recubierto por la resina (fig. 13.9d). Finalmente la fotorresina que queda se disuelve en soluciones ácidas calientes (fig. 13.9e). Sobre la superficie expuesta del silicio en la oblea se puede efectuar después bien sea la difusión de impurezas o bien la deposición de contactos, según el caso.

Una desventaja importante de este proceso de ataque con ácidos, denominado *grabado químico*, es que el ataque del óxido es isótropo, es decir, procede con la misma velocidad en todas las direcciones. Debido a ello, al transferir patrones de la máscara a la fotorresina existe también un ataque lateral del óxido que está debajo de la resina protectora y cercano a los bordes, con lo que se pierde resolución espacial. Este efecto se puede evitar con los modernos sistemas de *grabado en seco*, basados en el envío de un haz de iones, con energías de unos 500 eV, contra la superficie del silicio con el patrón de resina ya delineado. El campo eléctrico acelerador de los iones es perpendicular a la superficie de la oblea, por lo que no hay apenas bombardeo de iones en las paredes laterales de la máscara. Se consigue así un grabado prácticamente vertical, también denominado grabado anisótropo.

³ **Nota:** El término "resina" se utiliza frecuentemente como versión, no correcta, de la palabra inglesa "resist". Nosotros adoptamos aquí esta terminología por ser la utilizada más comunmente.

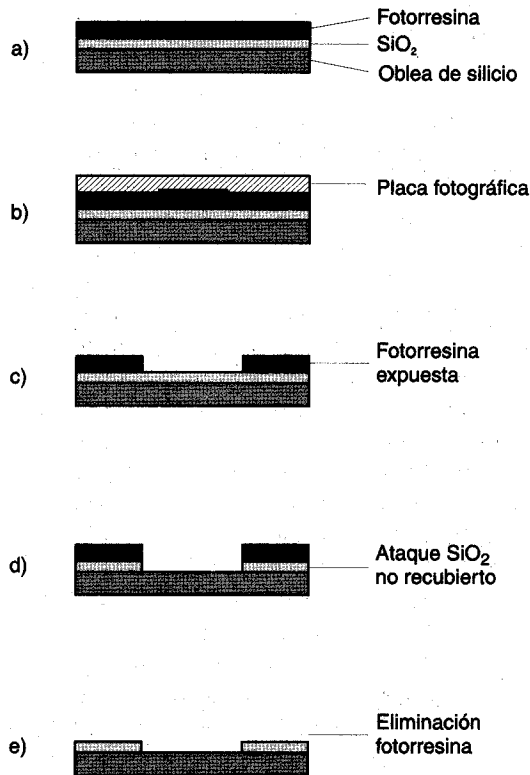


Fig. 13.9. Procesos fotolitográficos: a) Deposición de la fotorresina. b) Iluminación de la fotorresina por luz ultravioleta. c) Disolución de la fotorresina no expuesta a la radiación UV. d) Ataque químico del óxido de silicio. e) Disolución de la fotorresina remanente.

Los procesos litográficos con luz ultravioleta que hemos descrito tienen una resolución espacial limitada por la difracción de la radiación, la cual es aproximadamente igual a la longitud de onda de la radiación utilizada. Por tanto, para circuitos integrados submicrónicos (tecnología ULSI) a menudo es preciso recurrir a otro tipo de litografía, la *litografía por haz de electrones*. En este caso, se forma directamente la máscara con la resina depositada sobre el óxido. Para ello se barre con un haz de electrones la superficie de la resina, la cual se elige de forma que sea sensible a los electrones. La longitud de onda de los electrones utilizados se aproxima a tan sólo 1 \AA ; sin embargo, la resolución es mucho peor (del orden de $0.25 \text{ }\mu\text{m}$) debido a la dispersión lateral que sufren los electrones al incidir sobre la resina. Desde este

punto de vista, la *litografía por haz de iones* es la que presenta una mejor resolución, $0.1\ \mu\text{m}$, aproximadamente, aunque su introducción es tan reciente que apenas se utiliza a escala industrial. Una solución intermedia entre la litografía óptica y la basada en haz de iones es la provista por la *litografía de rayos X*. La longitud de onda de los rayos X, mucho menor que la de la radiación UV, hace que este tipo de litografía sea a veces utilizado en la tecnología ULSI. La litografía con rayos X hay que llevarla a cabo en vacío o en una atmósfera de helio ya que el aire absorbe fuertemente a los rayos X.

En la litografía de alta resolución hay que tomar una serie de precauciones, como son la utilización de placas fotográficas extremadamente planas, y el alineamiento perfecto entre las sucesivas placas utilizadas, ya que a menudo existen varias etapas de ataque durante la fabricación. De hecho, las ventanas abiertas en el óxido han de quedar situadas con un margen de error menor que la resolución alcanzada en el ataque. Es más, con objeto de aumentar el grado de precisión en todo el proceso de litografía, éste ha de realizarse enteramente en las denominadas *salas limpias*. Esto es debido a que la mayor parte de los fallos de los circuitos integrados se deben a partículas de polvo existentes en el ambiente, las cuales se depositan sobre las máscaras y la superficie del semiconductor actuando como un elemento opaco a la radiación. Por todo ello, la concentración de partículas de polvo por unidad de volumen debe ser cuidadosamente controlada mediante un filtrado apropiado del aire. En la fig. 13.10 se muestra la curva de distribución de los tamaños de partículas de polvo. Así, una sala limpia de clase 10 por ejemplo es la que presenta 10 partículas de polvo de tamaño $0.5\ \mu\text{m}$ o mayor por pie cúbico-

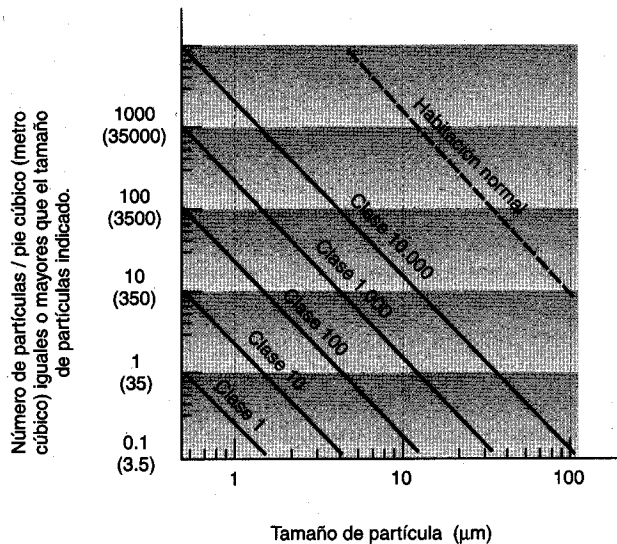


Fig. 13.10. Curva de distribución de los tamaños de las partículas existentes en la atmósfera de una sala limpia.

co (equivalente a una concentración de 350 partículas por metro cúbico). Para los procesos fotolitográficos se requiere al menos una sala limpia de clase 10, lo que significa que presenta, según se deduce de la fig. 13.10, una concentración de polvo unas 10^5 veces menor que la de una sala normal sin filtrado. De todos modos, afortunadamente, no todas las partículas de polvo que inciden sobre una superficie se adhieren a ella.

13.7. PROCESO DE DOPAJE DE UN SEMICONDUCTOR

El dopaje de los semiconductores se puede llevar a cabo bien por difusión a altas temperaturas a través del silicio de los dopantes provenientes de una fuente situada en la superficie o bien por implantación iónica. En este proceso, los iones de las impurezas se introducen en el semiconductor después de ser acelerados a grandes energías. Aunque el procedimiento únicamente empleado hasta la década de los 70 era el de difusión, en la actualidad se emplea tanto el proceso de difusión como el de implantación iónica.

13.7.1. Difusión de impurezas

En el proceso de difusión de impurezas en semiconductores, las impurezas se trasladan desde las regiones de alta concentración a regiones de concentración baja, de forma similar a lo que ocurre en la difusión de electrones y huecos en un semiconductor (apartado 2.5). El fenómeno de la difusión de impurezas tiene lugar por efecto de la temperatura, ya que a temperaturas ordinarias las impurezas del silicio, por ejemplo, son inmóviles para todos los efectos, mientras que a temperaturas elevadas las impurezas pueden alcanzar energía suficiente para moverse a través de la red cristalina. Tratándose de un proceso de difusión, se puede suponer en primera aproximación que el flujo j de partículas que atraviesan la unidad de área en la unidad de tiempo en la dirección x es proporcional a la variación de la concentración de impurezas, C , es decir:

$$j = -D \frac{\partial C}{\partial x} \quad [13.11]$$

Nótese la similitud de esta ecuación con las expresiones [2.33] y [2.34] para la corriente de difusión de electrones y huecos por efecto de la difusión. A la constante de proporcionalidad entre el flujo y el gradiente de la concentración se le denomina en este caso *coeficiente de difusión de las impurezas*, D , el cual depende la temperatura, T , según una ley de la forma:

$$D = D_0 \exp(-E_a/kT) \quad [13.12]$$

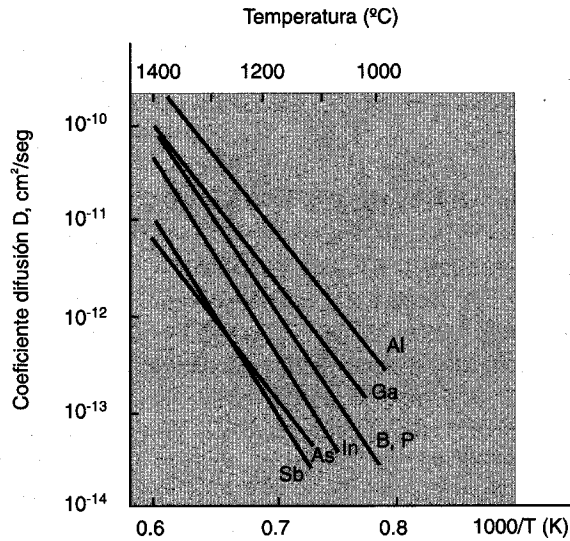


Fig. 13.11. Variación con el inverso de la temperatura del coeficiente de difusión de diversos dopantes en silicio.

donde E_a es la energía de activación para el proceso de difusión y D_0 es el valor del coeficiente de difusión extrapolado a temperatura infinita. En la fig. 13.11 se muestra en una escala logarítmica la variación con el inverso de la temperatura del coeficiente de difusión de los dopantes más importantes utilizados en el silicio. El comportamiento lineal del coeficiente D en esta gráfica prueba que se trata de un proceso activado térmicamente, de acuerdo con la ec. [13.12], siendo la pendiente de las rectas proporcional a la energía de activación E_a . En este proceso, los átomos dopantes se trasladan dentro del silicio en un proceso de difusión de vacantes, por lo que E_a comprende tanto la energía necesaria para trasladar un átomo de una posición vacante a otra como la energía que se requiere para crear las vacantes. El valor de E_a suele estar comprendido entre 3 y 5 eV.

Sustituyendo el valor del flujo de átomos dado por ec. [13.11] en la ecuación de continuidad para una dimensión, $\partial C / \partial t = - dj / dx$, y suponiendo que el coeficiente de difusión es independiente de la concentración de dopantes se obtiene la conocida *ley de Fick* para la difusión:

$$\frac{\partial C}{\partial t} = D \frac{\partial^2 C}{\partial x^2} \quad [13.13]$$

Resolviendo esta ecuación diferencial, junto con las condiciones de contorno, se puede obtener en cada caso la distribución de impurezas en el sólido durante el proceso de difusión.

Los procesos de difusión empleados en la fabricación de circuitos integrados suelen realizarse en dos etapas. En la primera, denominada *predifusión*, los átomos de impurezas procedentes de una fuente gaseosa son difundidos desde la superficie del semiconductor manteniendo constante en ésta la concentración de impurezas. A la predeposición sigue un proceso de *redistribución*, en el cual se corta la fuente que suministra las impurezas permitiendo la difusión hacia el interior del semiconductor. Debido a ello, en la difusión sólo intervienen las impurezas introducidas en la primera etapa que, por efecto de la temperatura, se distribuyen hasta alcanzar el perfil final de concentración. Entre el primero y segundo proceso se realiza un ataque químico a la superficie del semiconductor para remover el exceso de dopante en la superficie. En lo que sigue hallaremos las expresiones matemáticas de los perfiles de concentración resultantes de las dos etapas de difusión mencionadas.

- i) En la primera etapa, en la que la difusión de impurezas se realiza manteniendo una concentración superficial, C_s , constante durante el proceso, las condiciones de contorno para la variación de la concentración de impurezas en el interior del semiconductor, $C(x,t)$, son: $C(0,t) = C_s$ y $C(\infty,t) = 0$. Se puede demostrar que la solución de la ecuación de Fick [13.13] para estas condiciones de contorno es:

$$C(x,t) = C_s \operatorname{erfc} \left[\frac{x}{2(Dt)^{1/2}} \right] \quad [13.14]$$

cuya variación a distintos tiempos viene dada en la fig. 13.12a. El factor $(Dt)^{1/2}$ se denomina *longitud de difusión de las impurezas* (compárese con las ecs. [2.49] y [2.50]). La función complementaria de error de la expresión anterior, $\operatorname{erfc}(x)$, viene dada por:

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = 1 - \frac{2}{\pi^{1/2}} \int_0^x e^{-y^2} dy \quad [13.15]$$

La concentración de átomos de impureza, N , que quedan difundidos por unidad de área del semiconductor después de un cierto tiempo t , viene dado por:

$$N = \int_0^{\infty} C(x,t) dx \quad [13.16]$$

Sustituyendo la expresión de $C(x,t)$ dada por ecuación [13.14] y haciendo uso de las propiedades de la función $\operatorname{erfc}(x)$ se obtiene:

$$N = \frac{2}{\pi^{1/2}} C_s (Dt)^{1/2} \quad [13.17]$$

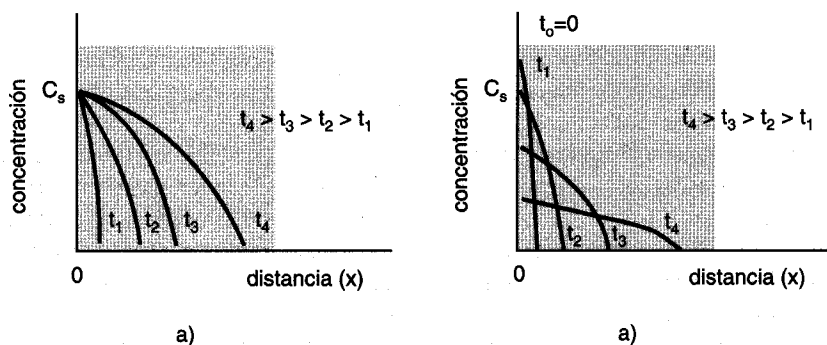


Fig.13.12. Concentración de impurezas en función de la distancia a la superficie del semiconductor cuando se mantiene la superficie: a) con una concentración constante de dopante durante el proceso de difusión y b) partiendo de una concentración fija, hasta que se agota el dopante.

la cual muestra que en esta etapa del proceso de difusión, la concentración total de impurezas aumenta con el tiempo siguiendo una ley potencial, con exponente 1/2.

Generalmente durante la etapa de predeposición las obleas se calientan a una temperatura previamente seleccionada, mientras se hace llegar a la superficie del semiconductor una concentración suficiente de átomos dopantes. Esto se consigue arrastrando hasta el reactor los gases o vapores procedentes de una fuente gaseosa o líquida que contiene los elementos o compuestos dopantes. Los átomos del material dopante penetran en el semiconductor hasta que se alcanza una máxima concentración, determinada por el límite de solubilidad. En la fig. 13.13 se muestran los límites de solubilidad a diferentes temperaturas de varias impurezas típicas del silicio.

- ii) En la segunda etapa, en la que la difusión se lleva a cabo utilizando una cantidad fija de dopante, con valor por unidad de área igual a N , las condiciones de contorno serán $C(\infty, t) = 0$, como antes, y aquella que expresa que la integral desde $x = 0$ hasta $x = \infty$ de la concentración $C(x, t)$ es igual a N . La solución de la ecuación de Fick con estas condiciones de contorno es:

$$C(x, t) = \frac{N}{(\pi Dt)^{1/2}} \exp\left(-\frac{x^2}{4Dt}\right) \quad [13.18]$$

que da una distribución gaussiana de las impurezas en el interior del semiconductor. Como es de esperar en este caso (véase fig. 13.12b), la concentración de dopante en la superficie C_s disminuye con el tiempo ya que si hacemos $x = 0$ en la ec. [13.18] se tiene:

$$C_s = \frac{N}{(\pi Dt)^{1/2}} \quad [13.19]$$

A la difusión en la que se parte de una cantidad fija de dopante en el interior del semiconductor también se la denomina *difusión con fuente instantánea* y se aplica en aquellos casos en que se requiere una concentración superficial relativamente baja simultáneamente con una gran profundidad de difusión. Este es el caso, por ejemplo, de la difusión requerida para la formación de la base de los transistores bipolares. Sin embargo, la concentración superficial que se obtiene en estos casos es demasiado baja para la formación del emisor, por lo que se suele recurrir a una difusión con *fente constante*.

En la fig. 13.14a se muestra un ejemplo del perfil de concentración resultante al difundir por ejemplo un dopante de tipo donador en un sustrato de tipo aceptor. En este caso, la concentración efectiva del dopante para cualquier valor de x viene dada por la diferencia entre las concentraciones de donores y aceptores. Así, por ejemplo, en el caso de un perfil de

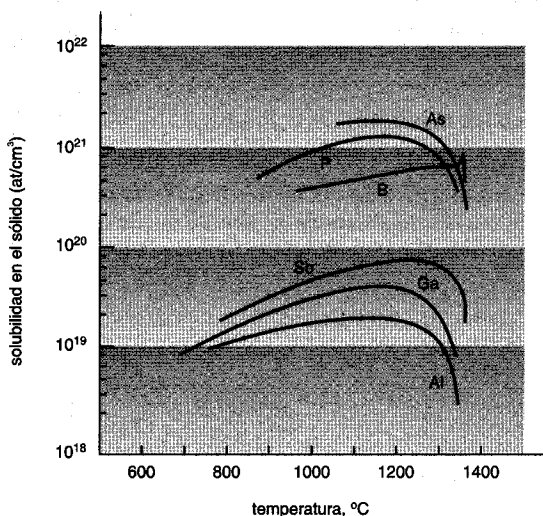


Fig. 13.13. Solubilidad de impurezas en silicio en función de la temperatura.

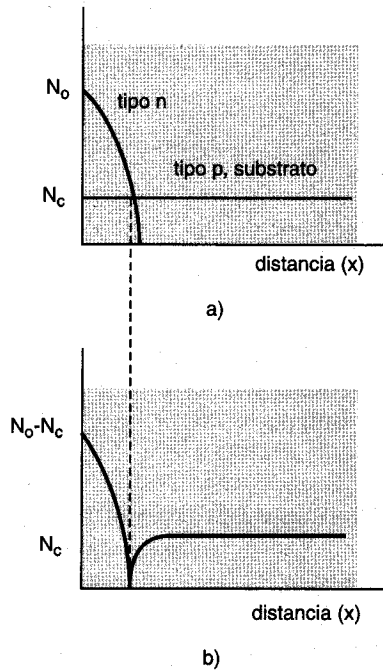


Fig. 13.14. Formación de una unión p-n por difusión: a) Perfiles de las impurezas de tipo n y de tipo p. b) Concentración neta de impurezas para la unión resultante.

impurezas de tipo n dado por la ec. [13.14], el substrato es de tipo n siempre que $C(x,t) > N_c$ donde N_c es la concentración del substrato. En cambio el substrato resulta de tipo p si $N_c > C(x,t)$. La fig. 13.14b muestra la concentración neta de impurezas de la unión p-n así formada. En esta figura, x indica la distancia de la unión a la superficie.

13.7.2. Implantación iónica

A diferencia de la difusión, la técnica de la implantación iónica consiste en un proceso de introducción de impurezas en semiconductores, realizado a baja temperatura. Las energías típicas de los iones de las impurezas incidentes sobre la muestra se suelen encontrar en el rango de 100 a 200 KeV y las dosis de iones implantadas típicamente varían entre 10^{12} y 10^{16} iones/cm². La ventaja de esta técnica reside en el hecho de que no es necesario calentar el semiconductor a temperaturas muy altas, con lo cual se elimina la posible redistribución de la

concentración de impurezas previamente establecida. También permite un control y reproducibilidad más precisos de la concentración de impurezas que en el caso de la difusión por temperatura. Otra propiedad reside en su flexibilidad al permitir por ejemplo la implantación de la base a través del emisor en transistores bipolares.

El esquema típico de un implantador se presenta en la fig. 13.15. En la fuente de iones se producen los iones de los átomos de impurezas de las especies dopantes, por ejemplo, B^+ y P^+ . Los iones se producen a partir de una descarga eléctrica a través del vapor de los átomos

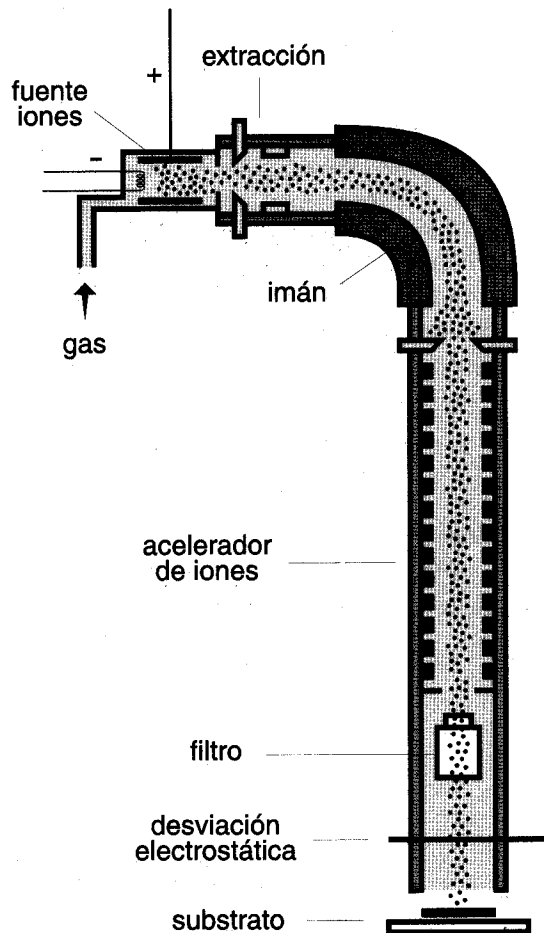


Fig. 13.15. Esquema de un implantador de impurezas en semiconductores.

que se pretenden ionizar. El haz de iones que se extrae de la fuente pasa luego por un imán que actúa como analizador de masas que elimina los iones no deseados. Los iones seleccionados son después acelerados a altas energías por un campo eléctrico y finalmente el haz de iones se dirige contra el sustrato que se desea dopar por implantación iónica.

Cuando los iones acelerados a altas energías penetran en un sustrato semiconductor comienzan a perder energía mediante interacciones con electrones y colisiones atómicas. Si la energía de los iones es muy alta predominan las interacciones electrónicas, pero a medida que decrece la energía las colisiones nucleares adquieren más importancia. En el caso de sustratos cristalinos la penetración de iones incidentes se hace mucho mayor cuando la dirección de incidencia coincide con una dirección cristalográfica ocurriendo el fenómeno denominado *canalaje* ("channeling"). Por ello lo normal es efectuar las implantaciones a unos 7° de una dirección cristalográfica principal, comportándose en este caso el sustrato prácticamente como si fuera amorfo.

Al penetrar un haz monoenergético de iones en un sustrato se produce una distribución estadística de los lugares en que se frenan los iones, después de recorrer caminos en zig-zag. El parámetro significativo es en este caso el denominado *rango proyectado*, R_p , que es la distancia que recorren los iones dentro del material en una dirección paralela al haz incidente. Otros parámetros que interesa conocer son las *fluctuaciones*, ΔR_p , (en inglés "straggle") en el valor de R_p , así como el valor de pico, N_p , de la concentración. La distribución de iones $C(x)$ a lo largo de la dirección de incidencia toma una forma aproximadamente gaussiana:

$$C(x) = N_p \exp \left[- \frac{(x - R_p)^2}{2\Delta R_p^2} \right] \quad [13.20]$$

donde x es la distancia recorrida en el interior del sustrato desde la superficie. La dosis total de iones por centímetro cuadrado, N_s , será entonces:

$$N_s = \int_0^{\infty} C(x) dx = N_p (2\pi)^{1/2} \Delta R_p \quad [13.21]$$

de donde:

$$N_p = \frac{N_s}{(2\pi)^{1/2} \Delta R_p} \quad [13.22]$$

Como se aprecia en esta ecuación, la concentración de pico, N_p , depende de la dosis total N_s y de la fluctuación típica en R_p .

En el caso de la litografía por haces de iones existen diferentes materiales que se pueden utilizar como máscaras. Para implantaciones de hasta unos 100 KeV se puede utilizar óxido de silicio, nitruro de silicio y fotorresinas. Sin embargo cuando se requieren implantaciones muy profundas, con energías de los iones incidentes del orden de MeV, se suele utilizar como máscaras un metal pesado (Au, Pt, etc.) sobre óxido de silicio.

13.8. METALIZACIONES

Al hablar de la fabricación de los dispositivos microelectrónicos en la sección 13.2 vimos que al final del proceso es necesario realizar contactos metálicos a los semiconductores. La mayor parte de las veces estos contactos son de tipo óhmico (apartado 4.2) excepto en los transistores MESFET (apartado 8.4), en los que el contacto es de tipo Schottky (apartado 4.1). Las metalizaciones también son necesarias cuando se une un dispositivo a otro mediante las denominadas pistas conductoras. En la reducción del tamaño de los circuitos integrados, las metalizaciones juegan un papel muy importante, ya que vienen a ocupar más de la mitad del espacio disponible para el circuito. Además, la mayor parte de los fallos en los circuitos integrados son debidos a defectos que surgen en las metalizaciones.

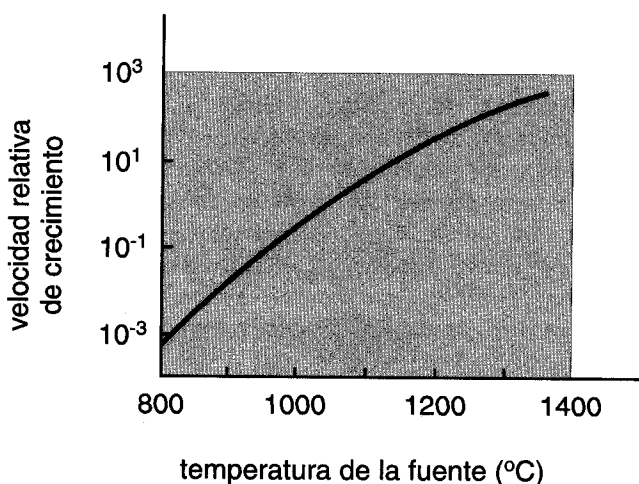


Fig. 13.16. Velocidad relativa de crecimiento de una capa delgada de oro en función de la temperatura del material de origen.

En todos estos casos, la metalización se lleva a cabo depositando una fina capa de metal sobre las regiones del sustrato donde se pretenden realizar los contactos o pistas conductoras. Existen diversos métodos para la formación de películas delgadas conductoras y en esta sección nos referiremos a los más empleados en la industria microelectrónica, como son los de evaporación, pulverización catódica ("sputtering") y deposición química en fase vapor (CVD). Los materiales más empleados son el aluminio y sus aleaciones, los metales refractarios (W, Ta, etc), el polisilicio (silicio dopado policristalino) y los siliciuros.

13.8.1. Evaporación térmica

La técnica de evaporación consiste en el calentamiento de un material (generalmente de tipo metálico) a una temperatura algo más alta que la de su punto de fusión, de modo que los átomos evaporados se condensan en el sustrato a recubrir formando una película metálica. Evidentemente, la evaporación se debe realizar en un aparato de vacío para evitar la oxidación del material. Se ha comprobado experimentalmente que el espesor de la capa depositada aumenta de modo casi exponencial con la temperatura del material de origen, según se aprecia en los resultados de la fig. 13.16 obtenidos para capas delgadas de oro.

Los diversos tipos de evaporación se diferencian unos de otros por el procedimiento de calentamiento del material a evaporar. El más sencillo -*evaporación por filamento incandescente*- consiste en la utilización de un filamento formado por un hilo de un metal de un punto de fusión muy alto, por ejemplo, tungsteno, al que se le da forma de arrollamiento y por el cual

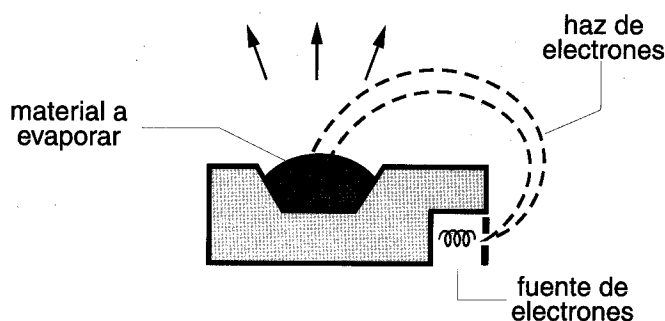


Fig. 13.17. Esquema de un sistema de evaporación por cañón de electrones.

se hace circular una corriente que produce calor por efecto Joule. En el interior del arrollamiento se coloca el metal a evaporar y se calienta el conjunto hasta que el material alcanza la temperatura de fusión haciendo pasar una corriente eléctrica. El material a evaporar, aún en estado fundido, se mantiene unido al filamento debido a la tensión superficial. En otros casos, y para evitar los problemas de adherencia del material fundido, el filamento se sustituye por una cinta metálica en forma de barquilla, también calentada por corriente eléctrica, sobre la cual se deposita el material a evaporar. En cualquiera de estos métodos, uno de los inconvenientes es que no permiten la evaporación de una cantidad elevada de material en un solo experimento. Existen también problemas de contaminación del sustrato con partículas del metal utilizado en el calentamiento (filamento o cinta metálica).

El método de calentamiento más utilizado en la industria microelectrónica es el de *bombardeo por cañón de electrones* (fig. 13.17). Según este método, los electrones producidos por un filamento incandescente mediante *emisión termoiónica* son acelerados mediante un voltaje elevado hasta alcanzar energías de unos 10 KeV. El haz de electrones es enfocado contra el blanco formado por el material a evaporar, y en la colisión con éste, los electrones pierden su energía cinética que se convierte en calor. Para evitar la contaminación del sustrato por el material que pudiera evaporarse del filamento termoiónico se utiliza a menudo una disposición del filamento y el blanco como la indicada en la fig. 13.17. Según se observa, los electrones realizan un giro de unos 270° por efecto de un campo magnético perpendicular a su trayectoria. Este método permite una velocidad de evaporación muy rápida, alrededor de 5000 Å por minuto.

13.8.2. Pulverización catódica

La pulverización catódica, a menudo denominada por su palabra inglesa "*sputtering*", consiste en el bombardeo de un *blanco o cátodo*, compuesto por el material a depositar, mediante un haz de iones, generalmente Ar^+ , generados mediante una descarga eléctrica entre un ánodo y el propio cátodo del material a evaporar. Durante el bombardeo del cátodo, los iones Ar^+ transfieren al blanco su momento cinético, por lo que los átomos de la superficie son eyectados para condensarse sobre el sustrato a recubrir. Lógicamente, el "*sputtering*" se realiza en un aparato de vacío, previamente evacuado antes de introducir el gas utilizado para la descarga. El gas debe ser inerte para que no reaccione con el blanco ni con el sustrato conforme se va depositando el material. Uno de los gases más utilizados es el argón debido a su bajo coste y a que los átomos del gas tienen una masa elevada. En la fig. 13.18 se muestra un equipo de "*sputtering*" típico, en el cual la descarga eléctrica se produce entre dos electrodos concéntricos: el ánodo en forma de placa circular y el cátodo con forma de anillo situado sobre él, en sus proximidades. Esta disposición suele ir provista de unos imanes situados sobre el cátodo, con objeto de confinar la descarga y aumentar la eficiencia del proceso ("*sputtering*" *magnetron*). Como consecuencia del bombardeo de los iones de Ar^+ , los átomos del cátodo son arrancados de la superficie y despedidos hacia el sustrato donde quedan depositados formando una película con una composición igual o similar a la del cátodo.

La técnica de "sputtering" presenta varias ventajas sobre la de evaporación térmica. Quizás la más importante es que el «sputtering» no requiere el calentamiento del blanco o del material a depositar a temperaturas altas. Es más, generalmente es preciso refrigerar el blanco para evitar que por efecto del bombardeo alcance temperaturas elevadas. El hecho de que los átomos del blanco sean despedidos como consecuencia del intercambio de momento con los

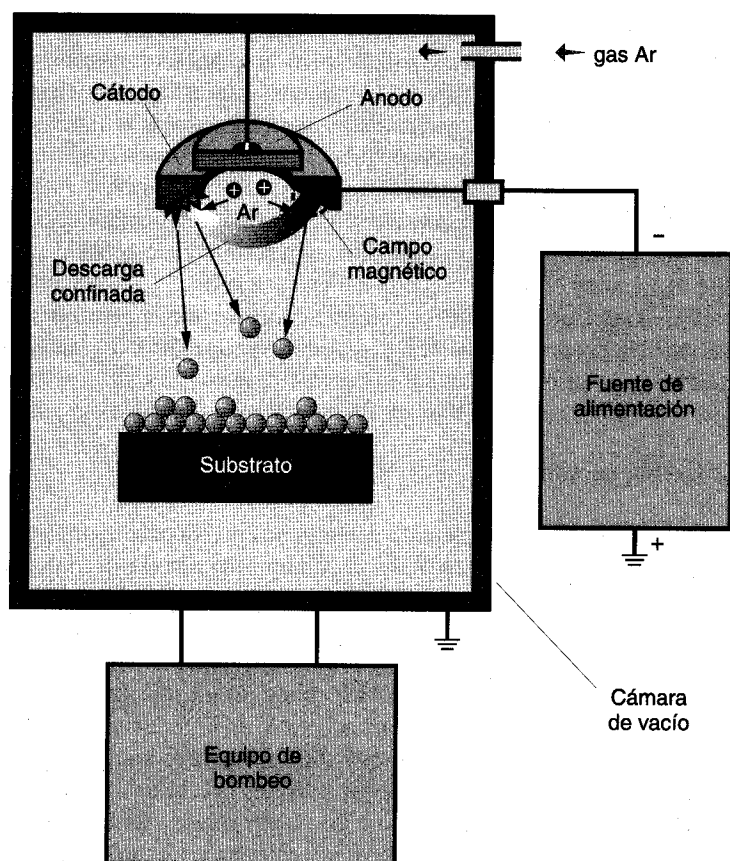


Fig. 13.18. Esquema de un equipo de "sputtering" magnetrón.

iones de Ar^+ , hace que sea posible depositar metales muy refractarios, es decir, de alto punto de fusión, tales como el tungsteno o el tantalio. Otra de las ventajas es que los átomos del material a depositar llegan al sustrato con bastante energía, por lo que la adherencia de las capas depositadas por "sputtering" es mayor que la que se obtiene por otras técnicas. En el caso de la deposición de aleaciones (la aleación Ni-Cr por ejemplo), la evaporación térmica presenta el inconveniente de que frecuentemente uno de los componentes (el Cr en este caso) se evapora mucho más fácilmente que el otro, dando lugar a que las capas depositadas tengan una composición diferente a medida que transcurre el proceso de deposición. Sin embargo, en el "sputtering" esto no ocurre y la capa depositada presenta la misma composición que el cátodo. Por último, el "sputtering" permite depositar películas en forma de compuesto del material del cátodo. Con este objeto se introduce con el argón gases reactivos como puede ser oxígeno o nitrógeno, los cuales permiten la formación de óxidos o nitruros, respectivamente ("sputtering" reactivo).

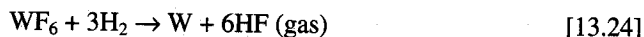
13.8.3. Deposición química en fase vapor

La deposición química en fase vapor o CVD ya ha sido expuesta en el apartado 13.5 en relación con la deposición de capas aislantes. Ahora bien, desde hace tan sólo unos pocos años esta técnica también se emplea en la deposición de capas metálicas. La técnica se basa en el empleo de compuestos volátiles, como haluros u organometálicos, que contengan los átomos del metal a depositar. A estos compuestos se les hace reaccionar sobre el sustrato situado en el interior de una cámara para dar, mediante calentamiento a temperaturas relativamente altas, el depósito metálico.

Aunque hay varios metales que se pueden depositar por CVD, la deposición del tungsteno a partir del hexafluoruro de tungsteno WF_6 es la más extendida. Cuando el tungsteno se deposita sobre silicio lo hace al principio de acuerdo con la reacción:



pero esta reacción cesa cuando la capa de W alcanza unos 200 Å de espesor, ya que el mismo W actúa como barrera de difusión entre el Si y el WF_6 . A partir de ahí la reacción prosigue por la reducción del WF_6 mediante hidrógeno según la reacción:



Otros metales como el cobre o el aluminio, también se pueden depositar mediante técnicas de CVD. En estos casos, se suele acudir a los compuestos organometálicos como gases precursores, ya que su descomposición se realiza a temperaturas más bajas.

13.8.4. Materiales utilizados en las metalizaciones

Al principio, el aluminio fue el metal más utilizado para contactos en semiconductores tanto en los transistores bipolares como en los MOS de los circuitos integrados. La ventaja del aluminio reside en su alta conductividad eléctrica y en la facilidad para formar contactos óhmicos con el silicio. Pero el aluminio presenta al mismo tiempo varios inconvenientes, uno de ellos derivado de su bajo punto de fusión y el otro de su capacidad de disolver al silicio.

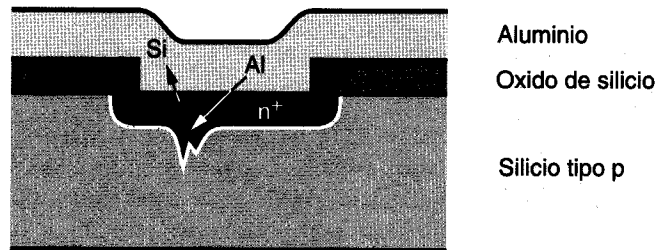


Fig. 13.19. Cortocircuito a través de una unión p-n producido por la disolución del silicio en el aluminio.

Así, los átomos de silicio disueltos producen espacios huecos que pueden ser rellenados por el mismo aluminio. De este modo y tal como se indica en la fig. 13.19 se pueden producir cortocircuitos en las uniones p-n. Una forma de solucionar parcialmente este problema consiste en depositar como contacto metálico aluminio previamente aleado con silicio hasta su saturación (aproximadamente el 1% de Si es suficiente) para así evitar la posterior disolución del silicio.

Se pueden formar contactos óhmicos de alta calidad sobre silicio utilizando como material de contacto ciertos compuestos de silicio con otro metal. Estos compuestos, denominados *siliciuros*, tienen una conductividad eléctrica cercana a la de los metales y forman con el silicio contactos de muy baja resistencia. Los siliciuros se forman frecuentemente depositando sobre el silicio una capa fina del metal, para hacerla reaccionar después con el propio

silicio mediante calentamiento. Los siliciuros más empleados pueden ser de metales nobles (PtSi, Pd₂Si) o de metales refractarios (TiSi₂, TaSi₂).

Los primeros transistores MOS, en la década de los 60, utilizaban el aluminio sobre el óxido de silicio como contacto de puerta. Sin embargo, algunos años más tarde el aluminio fue reemplazado por el *polisilicio* para disminuir el voltaje umbral (apartado 7.3) y porque el polisilicio puede soportar las altas temperaturas necesarias para el procesamiento de los transistores MOS. El polisilicio, denominado así por estar constituido por silicio policristalino, se suele depositar por CVD a partir del silano. Su conductividad es mucho más alta que la del silicio utilizado en dispositivos semiconductores, ya que está dopado con una concentración tan alta de impurezas (generalmente fósforo) que se forma un semiconductor de tipo degenerado (es decir, el nivel de Fermi llega a penetrar en la banda de conducción, véase sec. 2.2.2). Sin embargo, cuando se quiere conseguir una conductividad eléctrica más alta se debe sustituir el polisilicio por siliciuros, especialmente de metales refractarios. Se utiliza el polisilicio y los siliciuros en lugar de los metales porque éstos presentan problemas de estabilidad química, adherencia, etc. Algunos de estos problemas se han solucionado en los últimos años, y hoy día se emplean ya contactos de metales refractarios (W, Mo) en la tecnología ULSI.

Las pistas conductoras de los circuitos integrados transportan corrientes muy pequeñas. Sin embargo, como la sección transversal de ellas es de dimensiones muy reducidas, la densidad de corriente puede alcanzar valores de hasta 10^6 Acm^{-2} . Cuando la densidad de corriente es muy elevada se puede producir una transferencia de momento de los electrones de la corriente a los átomos de la pista conductora provocando que éstos se trasladen de un sitio a otro. En principio puede parecer sorprendente que los electrones, con una masa mucho menor que la de los átomos sean capaces de desplazar los átomos de la red de su posición de equilibrio. Sin embargo, hay que considerar que cuando se trata de densidades de corriente elevadas se produce una auténtica "lluvia" o bombardeo de electrones sobre los átomos del metal. De este modo se produce el fenómeno conocido con el nombre de *electromigración*. Mediante este mecanismo puede haber lugares donde el material es eliminado, lo cual llega a producir la interrupción del paso de la corriente. Por el contrario, en otros lugares puede haber una acumulación tal de material que da como resultado un cortocircuito entre dos pistas conductoras cercanas. Es evidente que el fenómeno de electromigración se verá muy aumentado por efecto de la temperatura. Se ha observado experimentalmente que, si en las pistas conductoras de Al-Si(1%) se añade un pequeño porcentaje de cobre, la electromigración queda muy disminuida. Esto es posiblemente debido a que el cobre se sitúa en las juntas de grano del material policristalino de la pista conductora disminuyendo de este modo la difusión de los átomos.

Cuando se hace contacto al semiconductor, con un siliciuro por ejemplo, y luego se efectúa el contacto del siliciuro con el aluminio de la pista conductora también se puede producir una alta electromigración entre estos dos materiales activada por la temperatura. Para

evitar este fenómeno es costumbre en los circuitos integrados hacer uso de las denominadas *barreras de difusión*, siendo las más conocidas las barreras de nitruro de titanio que impiden el paso de átomos a través de ellas y por otra parte presentan una resistencia eléctrica adicional muy baja.