Nuevas Tecnologías en los Dispositivos Electrónicos

Coordinador :
Francisco Gámiz Pérez
Autores:
Andrés Godoy Medina
Andrés Roldán Aranda
Carlos Sampedro Matarín
Juan Antonio Jiménez Tejada
Juan Bautista Roldán Aranda
Juan Enrique Carceller Beltrán
Francisco Gámiz Pérez
Francisco Jiménez Molinos
Pedro Cartujo Casinello

Dpto. Electrónica y Tecnología de Computadores
Universidad de Granada

AUTORES:	A. Godoy, A. Roldán, C. Sampedro, J.A. Jiménez Tejada, J.B. Roldán, J.E. Carceller, F.Gámiz, F. Jiménez, P. Cartujo Cassinello
TITULO:	NUEVAS TECNOLOGÍAS EN LOS DISPOSITIVOS ELECTRÓNICOS
Año publicación:	2008
ISBN:	978-84-691-4090-1
Depósito legal:	GR-1413-2008
Editorial:	Departamento de Electrónica y Tecnología de Computadores
Lugar publicación:	Granada

Resumen:
"Nuevas tecnologías en los dispositivos electrónicos" es un libro de texto cuyo principal objetivo es proporcionar los fundamentos básicos de los dispositivos electrónicos más comunes: uniones, el transistor bipolar y los transistores de efecto campo metal-óxido-semiconductor (MOSFET), metal-semiconductor (MESFET) y de unión (JFET). Antes de abordar estos dispositivos hay un primer capítulo donde se asientan las bases de los semiconductores. Después del estudio de estos dispositivos clásicos hay un capítulo dedicado a dispositivos optoelectrónicos. La tecnología de fabricación de circuitos integrados también se aborda en este libro con capítulos dedicados al crecimiento de semiconductores, su impurificación, el crecimiento controlado de otros materiales, la litografía y el grabado. A continuación se trata de forma específica la tecnología de fabricación de circuitos integrados y finalmente se tratan aspectos industriales de la fabricación de componentes. El libro termina con capítulos dedicados al modelado de dispositivos electrónicos y los procesos de fabricación. El libro forma parte de los resultados de un proyecto de innovación docente denominado "Aplicaciones de las nuevas tecnologías a
la enseñanza de los dispositivos electrónicos". Los contenidos de este libro también están disponibles en formato web en la página del departamento de "Electrónica y tecnología de computadores":
http://electronica.ugr.es/~amroldan/deyte/

Summary:
"New Technologies in electronic devices" is a textbook " intended for undergraduate courses. Our aim is to provide the basic principles of common semiconductor devices: junctions, the bipolar transistor, and field-effect-transistors (FET), such as the metal-oxide-semiconductor FET (MOSFET), the junction FET (JFET) and the metal-semiconductor FET (MESFET). In a first chapter, the fundamentals of semiconductors are given. After these classical devices are studied there is a chapter where optoelectronic devices are treated. There is a group of chapters devoted to silicon processing techniques such as oxidation, ion implantation, lithography, etching. The fabrication processes of integrated circuits are also described. The book finishes with two chapters devoted to modelling of electronic devices and processes. This textwook is one of the results of an educational project carried on at the University of Granada, called "Applications of the new technologies in teaching electronic devices". The contents of this book can also be browsed in the webpage
http://electronica.ugr.es/~amroldan/deyte/

Contenidos:
Fundamentos de semiconductores.
Uniones semiconductoras.
El transistor de unión bipolar.
La estructura metal aislante semiconductor.
El transistor de efecto de campo MOS.
El transistor de efecto de campo de unión.
El transistor de efecto campo metal-semiconductor.
Dispositivos optoelectrónicos.
Crecimiento de semiconductores.
Impurificación controlada de semiconductores.
Crecimiento y deposición de películas.
Litografía y grabado.
Tecnología de fabricación de circuitos integrados.
Aspectos industriales de la fabricación de componentes.
Modelos para simulación de dispositivos electrónicos.
Modelos para simulación de procesos.

Capítulo

1

Fundamentos de Semiconductores

Diagrama de bandas de un semiconductor tipo N

Índice

Objetivos

- Exponer la diferencia entre metales, aislantes y semiconductores.
- Presentar las principales características de los semiconductores y los tipos que existen.
- Introducir el concepto de hueco y explicar su contribución a la corriente en un semiconductor.
- Mostrar el modelo de diagramas de bandas para el estudio de los semiconductores.
- Exponer los fundamentos básicos de la estadística de semiconductores.
- Proporcionar las ecuaciones básicas para el cálculo de la densidad de portadores móviles en semiconductores.
- Estudiar las situaciones de desequilibrio y los procesos de generación y recombinación
- Analizar los mecanismos de conducción eléctrica: difusión y deriva.
- Exponer la ecuación de continuidad.

Palabras Clave

Semiconductor.
Electrón.
Hueco.
Enlace covalente.
Diagrama de bandas.
Impurezas donadoras y aceptadoras.
Nivel de Fermi.
Función de distribución.
Densidad de estados.
Densidad de portadores.
Generación.
Recombinación.
Equilibrio térmico.
Pseudoniveles de Fermi.
Corriente de arrastre.
Corriente de difusión.
Ecuación de continuidad.
Ecuación de difusión.
Relación de Einstein.
Movilidad.
Coeficiente de difusión.
Distribución de Fermi-Dirac.
Distribución de Maxwell-Boltzmann.
Semiconductor degenerado.

1.1 Introducción

En el presente tutorial estudiaremos dispositivos de estado sólido. Esto es, realizados sobre substratos sólidos (generalmente, silicio cristalino). Por tanto, no analizaremos válvulas de vacío ni, los recientemente propuestos, dispositivos moleculares.

A su vez, los sólidos se pueden clasificar según la ordenación espacial de los núcleos atómicos que lo constituyen. Como se observa en la figura, si éstos se ordenan de forma regular en el espacio se denominan sólidos cristalinos. Por el contrario, los sólidos amorfos presentan una disposición irregular de sus átomos. Finalmente, cuando la estructura regular no se conserva en todo en el espacio, sino en pequeñas regiones, decimos que el sólido es policristalino pues está formado por multitud de cristales.

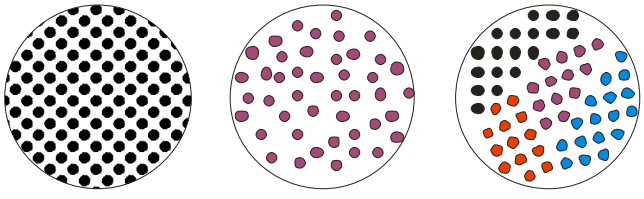


Figura 1.1.1 Esquema de la ordenación atómica en un sólido cristalino (a), amorfo (b) y policristalino (c)

La mayoría de los semiconductores de interés son sólidos cristalinos, por lo que nos centraremos en este tipo de sólidos. En este primer capítulo veremos cómo (y por qué) pueden ser conductores o no de la electricidad.

En el siguiente apartado analizaremos las principales diferencias entre metales, aislantes y semiconductores. Posteriormente nos centraremos en los semiconductores.

En primer lugar, estudiaremos cómo obtener el número de portadores de carga que pueden contribuir a la corriente eléctrica para poder conocer, de este modo, la conductividad de los semiconductores. Después veremos mediante qué procesos tiene lugar el transporte de los portadores de carga, es decir, qué mecanismos son responsables de la corriente eléctrica. También se estudiará que sucede al excitar (con luz, por ejemplo) un semiconductor.

1.2 Metales, aislantes y semiconductores

En esta sección veremos cuál es la diferencia, desde dos puntos de vista (fenomenológico y microscópico) que hay entre metales, aislantes y semiconductores. Después nos centraremos en la descripción de los semiconductores atendiendo a dos modelos: el del enlace covalente y el de las bandas de energía.

Descripción fenomenológica

A temperaturas próximas al cero absoluto (0 K) los metales son conductores de la electricidad, con una resistividad menor que ($\rho < 10^{-2}$ Ω cm). Por el contrario, los aislantes y los semiconductores no son conductores de la electricidad. Sin embargo, conforme se aumenta la temperatura:

- Los metales mantienen la conductividad casi constante, con un ligero descenso
- Los semiconductores incrementan su conductividad
- Los aislantes no conducen

Además, la conductividad de los semiconductores puede controlarse de forma muy precisa mediante procesos tecnológicos, como veremos a lo largo de esta lección.

Descripción según el tipo de enlace entre átomos

Como se sabe, existen tres tipos de enlaces entre átomos: iónico, metálico y covalente. A continuación recordamos brevemente cada uno de ellos haciendo hincapié especialmente en el característico de los semiconductores: el enlace covalente.

Enlace iónico

Se caracteriza porque los electrones están fuertemente ligados a los átomos y no pueden moverse a lo largo y ancho del sólido y no conducen, por tanto, la electricidad (aislantes). Un ejemplo es la sal (cloruro sódico), como se muestra en la Figura 1.2.1. Al constituirse el enlace, un electrón del átomo de sodio es atrapado por el átomo de cloro, de forma que se forman dos iones: uno positivo (formado por el núcleo del átomo de sodio y toda su nube electrónica excepto el electrón) y uno negativo (el átomo de cloro que ha atrapado un electrón). La atracción electrostática entre ambos átomos es la fuerza que los une para constituir el enlace. Obsérvese que en este tipo de enlace no se comparten los electrones, sino que éstos pertenecen a uno u otro átomo.

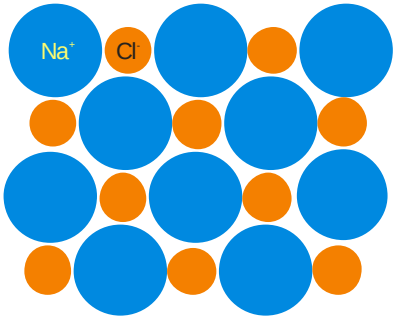


Figura 1.2.1 Enlace iónico

Enlace metálico

En los sólidos formados con este tipo de enlace, los electrones exteriores están desligados de los átomos, formando una nube electrónica distribuida en todo el sólido y que sirve de unión entre los núcleos atómicos. Por tanto, los electrones exteriores no están ligados a ningún átomo en concreto, y pueden moverse libremente bajo la acción de un campo eléctrico. Estos sólidos son, por consiguiente, buenos conductores de la electricidad.

Enlace covalente:

En este tipo de sólidos, los electrones de la capa más externa de cada átomo se comparten con los de otros átomos, formando el enlace entre ellos, de forma que cada par de electrones constituye un enlace entre átomos.

Por ejemplo, el silicio tiene cuatro electrones en su capa más externa y forma cuatro enlaces covalentes con otros tantos átomos de silicio (ver Figura 1.2.2).

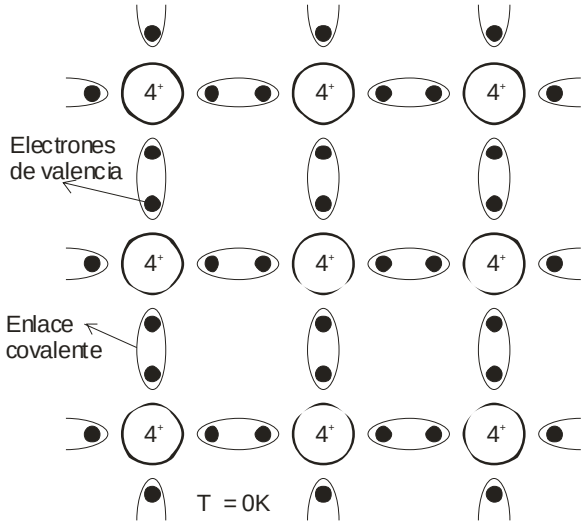


Figura 1.2.2 Esquema de un sólido covalente

En principio (cierto a T = 0 K), los electrones que forman el enlace se comparten por dos átomos y no pueden desplazarse por el cristal bajo la acción de un campo eléctrico. Por tanto, este material será aislante. Sin embargo, al aumentar la temperatura, la agitación térmica de los átomos de silicio puede provocar la ruptura de algunos enlaces, liberando electrones, como se indica en la Figura 1.2.3. Todos los electrones que han sido liberados pueden moverse por el cristal bajo la acción de un campo eléctrico, por lo que tenemos entonces que el sólido, que a bajas temperaturas es aislante, se comporta como un conductor de la electricidad a una temperatura lo suficientemente alta como para romper un número considerable de enlaces.

Semiconductores (modelo enlace covalente)

En este apartado realizaremos una descripción de los semiconductores que nos ayudará a entender cómo conducen la electricidad y cómo se puede modificar tecnológicamente su conductividad.

Concepto de hueco

En un semiconductor, la conducción se puede llevar a cabo de dos formas:

- Por los electrones libres
- Por los electrones ligados, que van ocupando sucesivamente diferentes enlaces

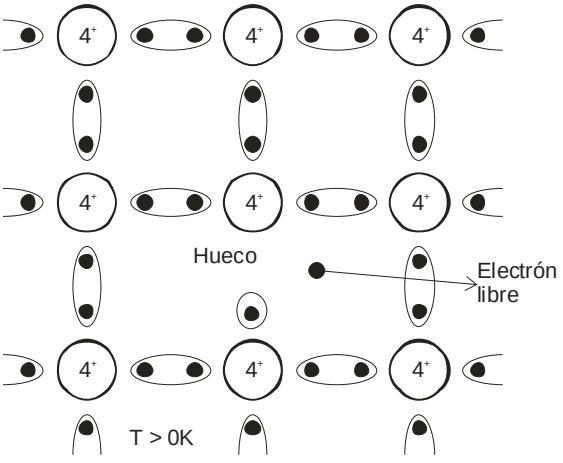


Figura 1.2.3 Creación de un electrón libre y de un hueco mediante la ruptura de un enlace covalente

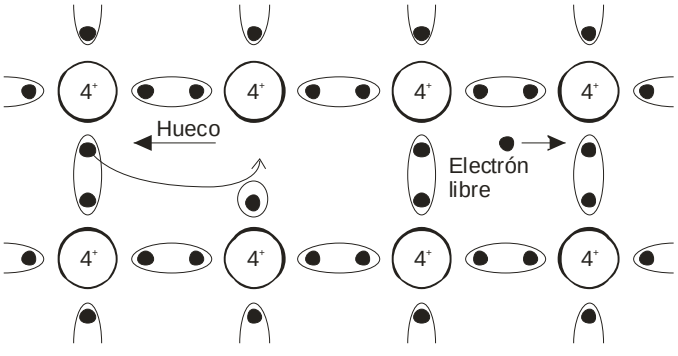


Figura 1.2.4 Movimiento de los electrones y de los huecos bajo un campo eléctrico

El movimiento de los electrones ligados equivale al de "una partícula" (hueco) de carga positiva y que se mueve en sentido contrario al que realmente llevan estos electrones.

Por tanto, considerando la contribución de ambos tipos de partículas, la conductividad viene dada por:

$$\sigma = q\mu_n n + q\mu_p p \,, \tag{1.1}$$

donde:

- n: concentración de electrones
- p: concentración de huecos
- μ_n: movilidad de los electrones
- μ_p: movilidad de los huecos
- q: valor absoluto de la carga del electrón

Semiconductores intrínsecos y dopados (tipo N y tipo P)

En principio, los huecos y electrones libres se crean por pares mediante la ruptura de enlaces covalentes. Por tanto, se cumple que:

$$n = p\ .$$

Esto es realmente así si el semiconductor no contiene impurezas que modifiquen la conductividad (aportando electrones o huecos, como veremos a continuación). En este caso, se dice que el semiconductor es intrínseco y a la concentración de electrones (o de huecos) se le denomina concentración intrínseca de portadores (se suele representar como n_i).
Por el contrario, los semiconductores dopados contienen impurezas para modificar de forma controlada la conductividad. Los átomos de las impurezas substituyen a los átomos del semiconductor (Si, por ejemplo), como se muestra en la figura. En general, en estos semiconductores se cumple que:

$$n \neq p.$$

Según las impurezas que introduzcamos, obtenemos dos tipos de semiconductores dopados (tipo N y tipo P), como describimos a continuación.

Dopado tipo N

Se obtiene por substitución de un átomo de silicio por algún átomo de la columna V de la tabla periódica (impurezas donadoras). En este caso, se aportan electrones libres para la conducción gracias a que se requiere poca energía para liberar uno de los electrones de la capa de valencia de cada átomo de impureza. El resto de los electrones del átomo forman enlaces covalentes con cuatro átomos de silicio vecinos. En conclusión, se obtiene un electrón disponible para conducir la electricidad y se queda un átomo fijo con carga positiva y con una configuración electrónica igual a la de un gas noble.

A temperatura ambiente y con una concentración de impurezas (N_D) suficientemente alta, la concentración de electrones será prácticamente igual a la concentración de impurezas donadoras ($n \sim N_D$), puesto que casi todas las impurezas habrán perdido su quinto electrón. Además, debido a que resulta más fácil romper el quinto enlace de la impureza que cualquiera de los formados entre átomos de silicio (que poseen la configuración de gas noble), la mayoría de los electrones disponibles para la conducción procederán de la ruptura de un enlace con una impureza y no de la ruptura de un enlace entre átomos de silicio (que, además, genera un hueco). Por tanto, se cumple que:

$$n \approx N_D >> p \Rightarrow \sigma \approx q\mu_n n \approx q\mu_n N_D\ . \qquad (1.2)$$

Como vemos en este caso, se cumple que la conductividad depende principalmente de la concentración de electrones, no de los huecos. Además, el número de electrones disponibles (n) está

determinado por la concentración de impurezas donadoras, que se puede fijar mediante procesos tecnológicos.

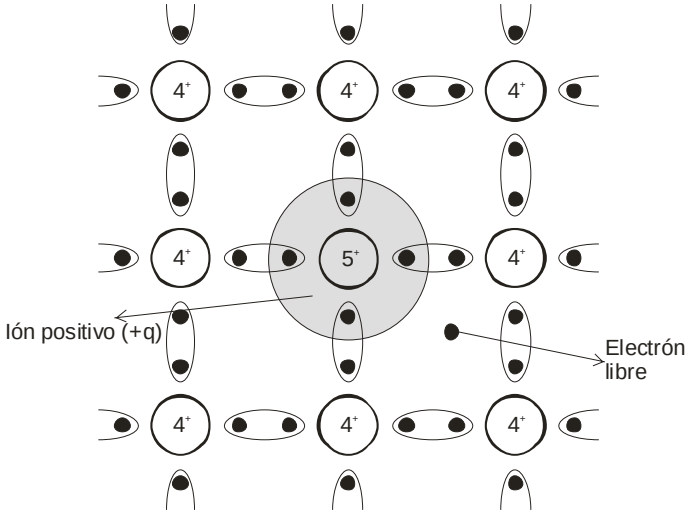


Figura 1.2.5 Semiconductor tipo N (con impurezas donadoras)

Dopado tipo P

Estos materiales se obtienen por la substitución de un átomo de silicio por algún átomo de la columna III de la tabla periódica (impurezas aceptadoras). Como estos átomos sólo tienen tres electrones en su capa más externa, formarán nada más que tres enlaces con átomos vecinos, quedando un enlace covalente no realizado. Se ha generado, por tanto, un hueco que puede moverse bajo la acción de un campo eléctrico, quedando entonces un átomo fijo con carga negativa.

Con una concentración de impurezas (N_A) suficientemente alta se cumple que $p \sim N_A >> n$. Por tanto, la conductividad depende principalmente por el movimiento de los huecos, cuyo número está determinado por la concentración de impurezas aceptadoras.

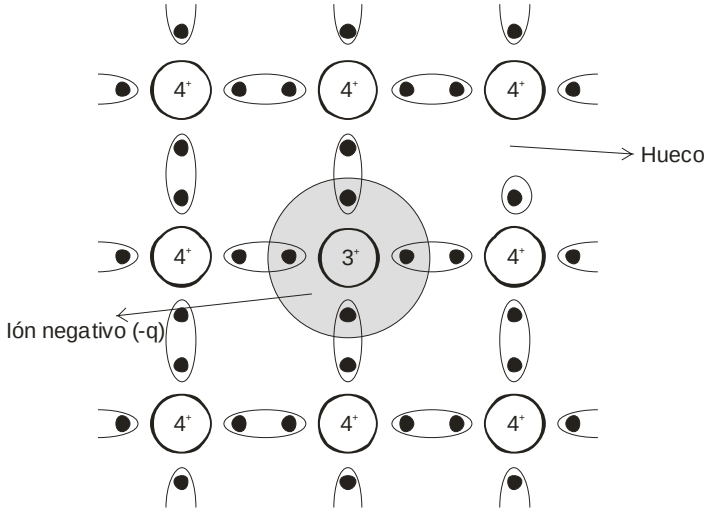


Figura 1.2.6 Semiconductor tipo P (con impurezas aceptadoras)

Modelo de bandas de energía

En este apartado vamos a describir una versión simplificada del modelo de bandas de energía basándonos en las características que ya conocemos de los semiconductores.

Introducción

El modelo de bandas de energía resulta muy útil para el estudio del movimiento de electrones y huecos en semiconductores. Como hemos dicho, en este curso veremos sólo un modelo simplificado, en el que se representan la energía de los electrones y de los huecos en función de la posición.

En este modelo, los electrones ligados (en un enlace covalente) se representan en una banda de energía llamada *banda de valencia*. Por otro lado, los electrones libres están situados en la *banda de conducción*. Ambas bandas están separadas por un rango de energías prohibidas, denominado *banda prohibida*.

A 0 K, la banda de valencia de un semiconductor está completa (no hay huecos) y no puede haber conducción porque tampoco hay electrones en la banda de conducción. Para que sea posible la conducción, deben generarse electrones (en la banda de conducción) y huecos en la banda de valencia. Para ello, un electrón de la banda de valencia debe

pasar a la de conducción. De hecho, el ancho de la banda prohibida es igual a la energía necesaria para romper un enlace covalente.

La diferencia que hay entre un semiconductor y un aislante es la anchura de la banda prohibida. En el aislante es tan grande que los electrones no pueden pasar de la banda de valencia a la de conducción con la única aportación de la energía térmica.

En el metal, la banda de conducción está medio llena incluso a bajas temperaturas y pueden conducir la electricidad.

Antes de seguir, señalaremos que la energía correspondiente al fondo de la banda de conducción se representará, en lo sucesivo, por E_c. Del mismo modo, E_v es la energía más alta de la banda de valencia.

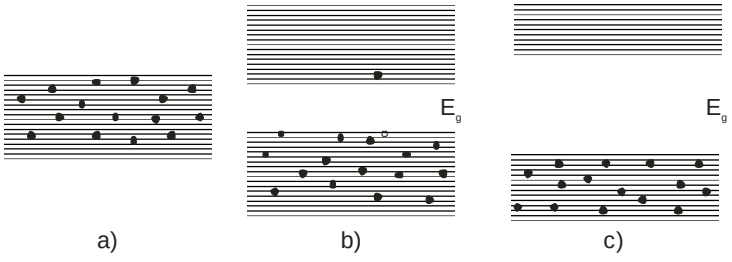


Figura 1.2.7 Diagrama de bandas de un metal (a), un semiconductor (b) y un aislante (c)

Semiconductores dopados en el modelo de bandas de energía

Semiconductores tipo N

La introducción de impurezas donadoras equivale a la presencia de un nivel energético en el interior de la banda prohibida, cercano al fondo de la banda de conducción. De este modo, los electrones situados en este nivel donador pueden saltar fácilmente a la banda de conducción, suministrándole electrones sin haber creado huecos.

Como la distancia entre ambas bandas es mucho mayor que la que hay entre la banda de conducción y el nivel donador que la que hay entre ambas bandas, es mucho más fácil que salte un electrón desde el nivel donador que desde la banda de valencia.

Semiconductores tipo P

La introducción de impurezas aceptadoras equivale a la presencia de un nivel energético en el interior de la banda prohibida, cercano a la banda de valencia, y que permite que se generen huecos al pasar electrones de la banda de valencia a este nivel aceptador.

1.3 Estadística de semiconductores

En este apartado daremos respuesta a las siguientes preguntas:

- ¿Cuántos electrones hay en la banda de conducción?
- ¿Cuántos huecos hay en la banda de valencia?
- ¿Cuántas impurezas están ionizadas?

Para ello:

- Primero veremos cuántos estados pueden ocupar los electrones
- Después aprenderemos cómo averiguar cuáles de ellos están realmente ocupados
- Finalmente, calcularemos el número de electrones y huecos disponibles para la conducción

Densidad de estados

No todas las energías están permitidas para los electrones y los huecos, sino que los electrones sólo pueden estar en ciertos *estados* con una energía determinada. La diferencia de energía entre estados dentro de una banda es muy pequeña, por lo que se puede hablar de una banda continua de estados con una densidad de estados por unidad de energía dada por:

$$g_c(E) = \lim_{\Delta E \to 0} \frac{\text{nº estados}}{\Delta E} . \qquad (1.3)$$

La densidad de estados no es uniforme en toda la banda de conducción, sino que depende de la energía (la separación entre estados es menor para niveles más separados de la banda prohibida). Se demuestra que:

$$g_c(E) = \frac{8\pi\sqrt{2}}{h^3}(m_e^*)^{3/2}\sqrt{E - E_c}, \quad \text{si } E > E_c , \qquad (1.4)$$

$$g_v(E) = \frac{8\pi\sqrt{2}}{h^3}(m_h^*)^{3/2}\sqrt{E_v - E}, \quad \text{si } E < E_v , \qquad (1.5)$$

donde:

$g_{c(E)}$: densidad de estados en la banda de conducción por unidad de energía

$g_{v(E)}$: densidad de estados en la banda de valencia por unidad de energía

E_c: mínimo de la banda de conducción

E_v: energía máxima de la banda de valencia

m_e: masa efectiva de los electrones para la densidad de estados

m_h: masa efectiva de los huecos para la densidad de estados

Ocupación de los estados en las bandas de conducción y valencia

La función de distribución de Fermi-Dirac

Los estados de las bandas de conducción y valencia no están uniformemente ocupados. En física estadística se demuestran las funciones que determinan la probabilidad de que un estado esté ocupado o no. Los electrones están controlados por la estadística de Fermi-Dirac, que establece que la probabilidad de que un cierto estado de energía E esté ocupado por un electrón viene dada por la siguiente función:

$$f(E,T) = \frac{1}{1 + e^{\frac{E - E_f}{kT}}}, \qquad (1.6)$$

donde:

- $f(E,T)$: función de distribución de Fermi-Dirac. Es la probabilidad de que un estado de energía E esté ocupado por un electrón. Toma valores entre 0 y 1
- E_f: nivel de Fermi o potencial electroquímico
- K: constante de Boltzmann

Como puede observarse, la función de distribución depende de la temperatura. En la Figura 1.3.1 se ha representado la función de Fermi-Dirac para tres valores de la temperatura. Con T = 0 K, la función de distribución es abrupta y se verifica:

$$\begin{cases} \text{si } E < E_F, & f(E, 0\,\text{K}) = 1 \\ \text{si } E > E_F, & f(E, 0\,\text{K}) = 0 \end{cases}$$

Conforme la temperatura aumenta, se ensancha la distribución (tanto más cuanto mayor sea T), indicando que estados de mayor energía pueden ser ocupados con mayor probabilidad que a temperaturas más bajas.

A continuación enumeramos algunas propiedades interesantes de la función de distribución de Fermi-Dirac:

- Para cualquier valor de la temperatura se cumple que para una energía igual al nivel de Fermi, la probabilidad de ocupación es 0.5:
$$\left[f(E = E_F, T) = \tfrac{1}{2} \right].$$
- La función $f(E,T)$ es simétrica respecto de E_F.

Por otra parte, si $f(E,T)$ es la probabilidad de que un estado de energía E esté ocupado por un electrón, entonces $(1 - f(E,T))$ es la probabilidad de que esté vacío y por tanto $(1 - f(E,T))$ proporciona la probabilidad de que un estado de la banda de valencia esté ocupado por un hueco.

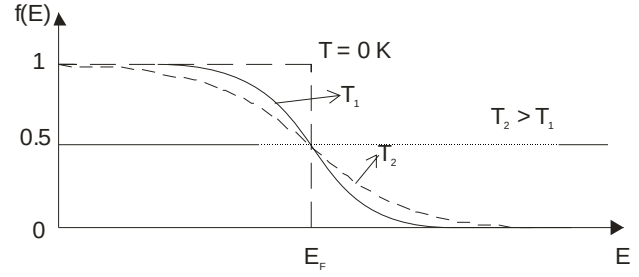


Figura 1.3.1 Función de Fermi-Dirac para varias temperaturas

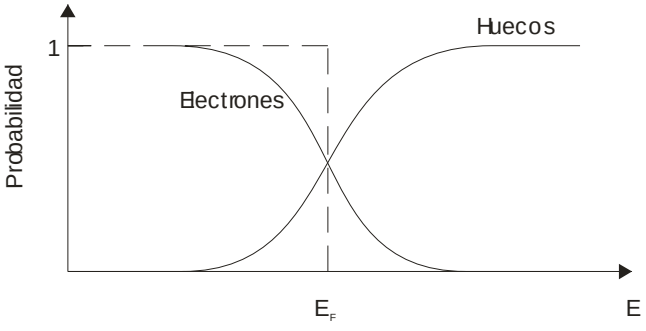


Figura 1.3.2 Función de distribución de Fermi-Dirac y probabilidad de que un estado esté ocupado o vacío

Finalmente, comentaremos que el nivel de Fermi es un potencial termodinámico, como la presión y la temperatura. Por tanto, el nivel de Fermi es constante en un sistema en equilibrio.

Ocupación de los estados en las bandas de conducción y valencia

El número de electrones con energía E es $g_c(E)f(E)$ y el de huecos con energía E $g_v(E)(1-f(E))$. Pero, ¿qué valor toma $f(E)$ en las bandas de conducción y valencia? La respuesta es que depende de la posición del nivel de Fermi. Analicemos a continuación dónde se sitúa el nivel de Fermi en los distintos tipos de semiconductores.

Como se ilustra en la Figura 1.3.3:

a) En un semiconductor intrínseco el nivel de Fermi está aproximadamente en la mitad de la banda prohibida (por simetría de la función de Fermi-Dirac), puesto que tiene que haber el mismo número de electrones y de huecos. Por tanto:

$$E_F \equiv E_i \approx \frac{E_c+E_v}{2}.$$

Hemos denominado a la posición del nivel de Fermi en el semiconductor intrínseco como E_i.

b) En un semiconductor tipo N, hay muchos más electrones que huecos, por lo que E_F debe estar más cerca de la banda de conducción que de la de valencia.

c) Finalmente, en un semiconductor tipo P, E_F está próximo a la banda de valencia para que haya mayor número de huecos que de electrones.

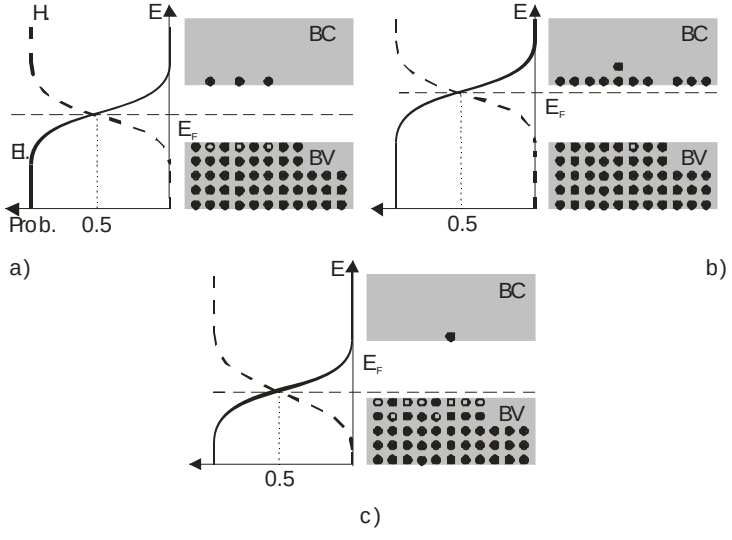


Figura 1.3.3 Posición del nivel de Fermi en un semiconductor intrínseco (a), tipo N (b) y tipo P (c)

Concentración de portadores en equilibrio térmico

Una vez que en los apartados anteriores hemos determinado la densidad de estados y la probabilidad de que estén ocupados, estamos ya en condiciones de determinar la magnitud que nos interesa: la concentración de electrones y huecos disponibles para la conducción.

La concentración de electrones en la banda de conducción viene dada por:

$$n_0 = \int_{E_c}^{\infty} g_c(E) f(E) dE \,, \tag{1.7}$$

mientras que la de huecos en la banda de valencia es:

$$p_0 = \int_{-\infty}^{E_v} g_v(E)\big(1 - f(E)\big) dE \,. \tag{1.8}$$

El subíndice 0 indica que estas expresiones son válidas en situación de equilibrio térmico. Substituyendo las expresiones de las densidades de estado y de la función de Fermi-Dirac que hemos mostrado en los apartados anteriores podemos calcular estas integrales.

Desafortunadamente, no se obtiene una solución analítica general. Sin embargo, podemos realizar aproximar la función de Fermi-Dirac por:

$$f(E) \approx e^{-\frac{E-E_f}{kT}} . \tag{1.9}$$

Esta aproximación es bastante buena si se cumple que:

$$\left|E_{c,v} - E_F\right| >> kT .$$

Cuando esto se verifica, se dice que el semiconductor es no degenerado y que cumple la función de distribución de Maxwell-Boltzmann (la función de la ecuación (1.9), que es la función de distribución correspondiente a partículas clásicas.

En este caso particular (bastante común) se obtienen las siguientes concentraciones electrones y huecos:

$$\begin{aligned} n_0 &= N_c(T)e^{-\frac{E_c-E_F}{kT}} \\ p_0 &= N_v(T)e^{\frac{E_v-E_F}{kT}} \end{aligned} . \tag{1.10}$$

Obsérvese que podemos expresar la densidad de electrones como $N_C(T)f(E_C)$ y la de huecos como $N_V(T)(1\text{-}f(E_C))$,por lo que se puede identificar $N_c(T)$ con una densidad efectiva de estados correspondiente a la banda de conducción. Es decir, es como si hubiésemos reducido toda la banda a la energía E_c y todos los posibles estados de la banda de conducción estuviesen condensados en el fondo de la banda de conducción. Análogamente, se habla de la densidad efectiva de estados correspondiente a la banda de valencia ($N_v(T)$). Estas densidades de estados vienen dadas por:

$$N_c(T) = 2\left(\frac{2\pi m_n^* kT}{h^2}\right)^{3/2} , \tag{1.11}$$

$$N_v(T) = 2\left(\frac{2\pi m_h^* kT}{h^2}\right)^{3/2} . \tag{1.12}$$

Veamos ahora cuál es la relación que hay entre las concentraciones de electrones y huecos. Calculemos en primer lugar el valor de su producto:

$$n_0 p_0 = N_c N_v e^{\frac{E_v-E_c}{kT}} = N_c N_v e^{\frac{-E_g}{kT}} . \tag{1.13}$$

En el caso particular de un semiconductor intrínseco, se cumple que $n_i = p_i$ y que $n_i p_i = n^2_i$. Por tanto:

$$n_i = p_i = \sqrt{N_c N_v}\, e^{-\frac{E_g}{2kT}} . \tag{1.14}$$

La concentración intrínseca de portadores en un semiconductor viene determinada por la anchura de la banda prohibida. Además, para

cualquier tipo de semiconductor, extrínseco o intrínseco, podemos expresar la ecuación (1.13) como:

$$n_0 p_0 = n_i^2 \qquad \text{(Ley de acción de masas)}. \tag{1.15}$$

Vemos que cuanto mayor sea la concentración de un tipo de portadores, menor es la concentración del otro tipo ($n_0 = n_i^2/p_0$). Por eso, en un semiconductor extrínseco se habla de portadores *mayoritarios* y *minoritarios*.

Por otro lado, otra forma conveniente de expresar las concentraciones de portadores es referirlas a las concentraciones intrínsecas:

$$\begin{aligned} n_0 &= n_i e^{\frac{EF-E_i}{kT}} \\ p_0 &= n_i e^{\frac{E_i-E_f}{kT}} \end{aligned}. \tag{1.16}$$

Volvemos a recalcar que estas ecuaciones son *válidas sólo en equilibrio térmico* (ausencia de campo externo).

Cálculo práctico de las concentraciones de portadores

Para calcular las concentraciones de portadores en semiconductores extrínsecos con las expresiones del apartado anterior hay que conocer la posición del nivel de Fermi. No obstante, usualmente se realizan algunas aproximaciones que lo evitan.

Semiconductores intrínsecos

¿Cuál es la posición del nivel de Fermi en un semiconductor intrínseco (E_i)? La respuesta se obtiene igualando n_i y p_i (ecuaciones (1.10)) y despejando:

$$E_i = \frac{E_c + E_v}{2} + \frac{kT}{2} \ln \frac{N_v}{N_c}. \tag{1.17}$$

Como puede verse, quedaría justo en la mitad de la banda prohibida si las densidades efectivas de estados en la banda de valencia y de conducción fuesen las mismas.

Semiconductores extrínsecos

Tipo N

Para obtener las densidades de portadores en un semiconductor extrínseco tipo N, en principio deberíamos emplear además de (1.10) ó (1.16) la siguiente ecuación de neutralidad (puesto que desconocemos la posición del nivel de Fermi):

$$n_0 = p_0 + N_D^+ . \tag{1.18}$$

Donde la concentración de impurezas ionizadas viene dada por:

$$\frac{N_D^+}{N_D} = 1 - \frac{N_D^0}{N_D} = \frac{1}{1 + e^{\frac{E_F - E_D}{kT}}} . \tag{1.19}$$

Con (1.10), (1.18) y (1.19) tenemos tres ecuaciones y tres incógnitas (n_0, N_D^+ y E_F). Sin embargo, estos cálculos resultan bastante laboriosos y normalmente se realizan las siguientes aproximaciones:

- A las temperaturas de interés todas las impurezas están ionizadas. Esto implica que:

$$N_D^+ \approx N_D .$$

- En la mayoría de casos prácticos se cumple que $N_D >> n_i$, por lo que:

$$n_0 >> p_0 .$$

Con estas aproximaciones se pueden calcular fácilmente las concentraciones de portadores y la posición del nivel de Fermi:

$$\begin{aligned} n_0 &\approx N_D \\ p_0 &\approx \frac{n_i^2}{N_D} \\ E_c - E_F &= kT \ln \frac{N_c}{N_D} \end{aligned} \tag{1.20}$$

Tipo P

Análogamente, para un semiconductor tipo P habría que plantear la siguiente ecuación de neutralidad:

$$p_0 = n_0 + N_A^- , \tag{1.21}$$

con:

$$\frac{N_A^-}{N_A} = \frac{1}{1 + e^{\frac{E_A - E_F}{kT}}} . \tag{1.22}$$

En este caso también se suelen realizar las siguientes aproximaciones:

- A las temperaturas de interés todas las impurezas están ionizadas:

$$N_A^- \approx N_A .$$

- En la mayoría de casos prácticos se cumple:

$$N_A >> n_i \Rightarrow p_0 >> n_0 .$$

Por lo que se verifica entonces que:

$$\begin{aligned} p_0 &\approx N_A \\ n_0 &\approx \frac{n_i^2}{N_A} \\ E_F - E_v &= kT \ln \frac{N_v}{N_A} \end{aligned} \qquad (1.23)$$

Semiconductores compensados

Los semiconductores compensados poseen tanto impurezas donadoras como aceptadoras.

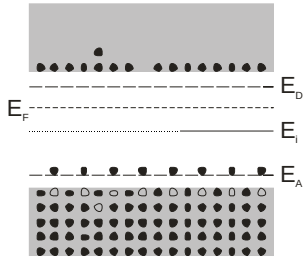


Figura 1.3.4 Diagrama de bandas de un semiconductor compensado

Desde el punto de vista de la concentración de portadores, se comportan como semiconductores tipo N (o tipo P) con:

$$N_D \rightarrow N_D - N_A \quad (\text{ó} \quad N_A \rightarrow N_A - N_D \ ).$$

La ecuación de neutralidad correspondiente es:

$$n_0 + N_A^{=} = p_0 + N_D^{+}. \qquad (1.24)$$

Además, la posición del nivel de Fermi debe verificar (1.10), (1.19), (1.22)y (1.24), por lo que habría que resolver este sistema de ecuaciones. Sin embargo, en este caso, también se supone que todas las impurezas están ionizadas, cumpliéndose entonces que:

$$n_0 \approx N_D - N_A$$

en un semiconductor tipo N, ó

$$p_0 \approx N_A - N_D$$

en un tipo P).

1.4 Concentraciones de portadores en desequilibrio. Pseudoniveles de Fermi

Equilibrio térmico. Generación y recombinación

Equilibrio térmico

Hasta ahora, todos nuestros cálculos han supuesto una condición de equilibrio térmico. De hecho, las expresiones anteriores (para el cálculo de la concentración de portadores) y el concepto de nivel de Fermi son sólo válidos en esta situación.

En equilibrio los electrones y huecos se generan por excitación térmica. Además, también existe un proceso de aniquilación de los electrones y huecos. Ambos procesos están balanceados de forma que las concentraciones de electrones y huecos se mantienen constantes (y vienen dadas por la estadística de Fermi-Dirac). En semiconductores intrínsecos, los electrones y huecos se crean por parejas y se cumple (en equilibrio térmico):

$$g_0 = r_0 , \tag{1.25}$$

donde:

- g_0: número de pares electrón-hueco generados por unidad de tiempo
- r_0: número de pares electrón-hueco recombinados por unidad de tiempo

Si aumenta g_0 (p.e., al aumentar la temperatura), también aumenta la tasa de recombinación para alcanzar una nueva situación de equilibrio (ec. (1.25)).

Generación y recombinación

Por su parte, la tasa de recombinación de pares e-h (r) verifica:

$$r = \alpha_r np \quad (r_0 = \alpha_r n_0 p_0 = a_r n_i^2 , \text{en equilibrio}), . \tag{1.26}$$

Es decir, la tasa de recombinación es proporcional a la concentración de electrones y huecos (como es lógico suponer, cuantos más haya, más probabilidad habrá de que se encuentren y se recombinen). α_r es un cierto coeficiente de proporcionalidad. Además, podemos conocer la tasa de generación térmica de pares *e-h* a partir de la ecuación de equilibrio (1.25):

$$g = a_r n_i^2 . \tag{1.27}$$

En general, esta igualdad es cierta incluso fuera del equilibrio.

Desequilibrio

Portadores en desequilibrio

Las concentraciones de portadores correspondientes al equilibrio térmico se pueden modificar mediante la acción de un agente exterior. Consideremos el caso de un semiconductor iluminado. El número de pares electrón-hueco creados, por unidad de tiempo, es igual a la tasa de generación menos un término debido a la recombinación:

$$\frac{dn}{dt} = \frac{dp}{dt} = g - r = \left(g_0 + G\right) - \alpha_r np\,, \tag{1.28}$$

donde G es la tasa de generación provocada por la iluminación del semiconductor.

Si dejamos transcurrir el tiempo se llega a una situación estacionaria, en la que se generan y recombinan los mismos pares electrón-hueco por unidad de tiempo:

$$\frac{dn}{dt} = \frac{dp}{dt} = 0 \Rightarrow \left(g_0 + G\right) = \alpha_r np\,. \tag{1.29}$$

Supongamos que ahora repentinamente cesa la acción externa. ¿Qué sucede? El semiconductor intenta recuperar la situación de equilibrio térmico mediante procesos de generación-recombinación, de forma que en todo instante se verifica:

$$\frac{dn}{dt} = \frac{dp}{dt} = g - r = g_0 - \alpha_r np = \alpha_r n_i^2 - \alpha_r np = \alpha_r \left(n_i^2 - np\right). \tag{1.30}$$

Esto es, se cumple que:

Si hay un exceso de portadores $(np > n_i^2) \Rightarrow \frac{dn}{dt} < 0$ (domina la recombinación)

Si hay un defecto de portadores $(np < n_i^2) \Rightarrow \frac{dn}{dt} > 0$ (domina la generación)

La cuestión que nos planteamos ahora es: ¿cuánto tarda en alcanzarse el equilibrio? La solución se obtiene al resolver la ecuación diferencial (1.30). En general, es no lineal y difícil de resolver. Sin embargo, tiene sencilla solución en el siguiente caso particular, correspondiente a un bajo nivel de inyección (esto es, el desequilibrio no es muy pronunciado: se han generado pocos pares electrón-hueco).

Recuperación del equilibrio con baja inyección

Desarrollemos la ecuación (1.30):

$$\begin{aligned} \frac{dn}{dt} &= \frac{d(\delta n)}{dt} = \alpha_r \left(n_0 p_0 - (n_0 + \delta n)\cdot(p_0 + \delta p)\right) = \\ &= -\alpha_r \left(n_0 \delta p + p_0 \delta n + \delta n \delta p\right) \approx -\alpha_r \left(n_0 \delta p + p_0 \delta n\right). \end{aligned} \tag{1.31}$$

En este caso se cumple:

$$\delta p = \delta n$$

y por tanto:

$$\frac{d(\delta n)}{dt} = -\alpha_r (n_0 + p_0)\delta n \,. \qquad (1.32)$$

Agrupando términos e integrando se resuelve la ecuación diferencial (1.32):

$$\delta n(t) = \Delta n e^{-t/\tau} \,. \qquad (1.33)$$

donde:

- Δn: exceso de portadores tras el cese de la iluminación (en $t = 0$)
- τ es el tiempo medio de vida de los portadores de carga en desequilibrio o constante de tiempo de recombinación y viene dado por:

$$\tau = 1/\alpha_r (n_0 + p_0) \,.$$

El caso particular de un semiconductor extrínseco (por ejemplo, tipo *N*) resulta de interés porque se verifica que:

$$\delta n \approx n,$$
$$p_0 \approx n_0$$

y, por tanto, se cumple que:

$$\begin{aligned} n(t) &= n_0 + \delta n \approx n_0 \\ \frac{dp(t)}{dt} &= -\frac{\delta p(t)}{\tau_p} \quad \text{con} \quad \tau_p = (\alpha_r n_0)^{-1} \\ p(t) &= p_0 + \delta p = p_0 + \Delta p e^{-t/\tau_p} \end{aligned} \qquad (1.34)$$

En este caso, el tiempo de vida medio se ha nombrado como τ_p y se le suele llamar *tiempo de vida medio de los portadores minoritarios.*

Pseudoniveles de Fermi

En este apartado vamos a introducir el concepto de pseudonivel de Fermi. Ya sabemos que la acción de un agente exterior (campo eléctrico, inyección de portadores, iluminación, ...) puede variar la concentración de portadores correspondiente al equilibrio térmico. Si la acción es constante en el tiempo se llega a una situación estacionaria ($g = r$), con unas concentraciones de portadores diferentes de las del equilibrio pero independientes del tiempo:

$$\begin{aligned} n &= n_0 + \delta n \\ p &= p_0 + \delta p \end{aligned} \,. \qquad (1.35)$$

Como ya hemos indicado anteriormente, estas concentraciones de portadores ya no vienen descritas por la estadística de Fermi-Dirac y no pueden calcularse aplicando las fórmulas anteriormente vistas. Para ilustrar este hecho, consideremos el siguiente ejemplo:

Ejemplo

Supongamos una muestra semiconductora que se ilumina uniformemente con luz de energía superior al gap. Se generan pares electrón-hueco, de forma que aumenta la concentración tanto de huecos como de electrones. De acuerdo con la estadística de Fermi-Dirac:

$$\left.\begin{array}{l} \text{- un aumento de } n \Rightarrow E_F \uparrow \\ \text{- un aumento de } p \Rightarrow E_F \downarrow \end{array}\right\} \Rightarrow$$

$\Rightarrow \nexists E_F$ que describa la concentraciones n y p simultáneamente.

Como vemos en este ejemplo, en un sistema en desequilibrio no podemos hablar de nivel de Fermi. En estos casos, se definen entonces los pseudoniveles de Fermi, E_{FN} y E_{FP}, de forma que se verifica:

$$\begin{aligned} n &= N_c e^{-\frac{E_c - E_{Fn}}{kT}} \\ p &= N_v e^{\frac{E_v - E_{Fp}}{kT}} \end{aligned} \qquad (1.36)$$

En general, los pseudoniveles de Fermi son distintos ($E_{FN} \neq E_{FP}$) aunque sus valores no son independientes (en el caso del ejemplo tenemos la ligadura $\delta n = \delta p$). Además se cumple:

$$np = N_c N_v e^{-\frac{E_g}{kT}} e^{\frac{E_{Fn} - E_{Fp}}{kT}} \Rightarrow np = n_i^2 e^{\frac{eV_{np}}{kT}} \qquad (1.37)$$

Como vemos, la concentración de portadores difiere tanto más de la correspondiente al equilibrio ($n_0 p_0$) cuanto mayor sea la separación de los pseudoniveles (eV_{np}). Además:

$$E_{Fn} > E_{Fp} \;\; (V_{np} > 0) \Rightarrow np > n_i^2 \Rightarrow \text{exceso de portadores}$$

$$E_{Fn} < E_{Fp} \;\; (V_{np} < 0) \Rightarrow np < n_i^2 \Rightarrow \text{defecto de portadores.}$$

Nota importante: los pseudoniveles de Fermi pueden ser dependientes de la posición, al contrario que el nivel de Fermi, que define un sistema en equilibrio y es constante.

1.5 Transporte de carga. Corriente en semiconductores

La corriente eléctrica en un semiconductor tiene dos contribuciones que analizaremos a continuación: la corriente de arrastre y la de difusión.

Corriente de arrastre

La corriente de arrastre se debe a la presencia de un campo eléctrico que acelera a los electrones y huecos. Se denomina también corriente de deriva (*drift* en inglés). En ausencia de campo eléctrico, los portadores se mueven aleatoriamente y no hay un transporte neto de carga. Con la presencia de un campo eléctrico, al movimiento aleatorio se superpone una componente (la misma para todos los portadores de igual carga) que provoca un movimiento neto en la dirección del campo eléctrico.

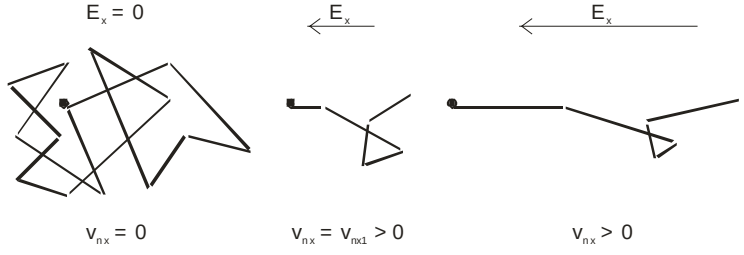


Figura 1.5.1 Movimiento aleatorio térmico de un electrón (a) y bajo la acción de un campo eléctrico (b)

El resultado es una velocidad media en la dirección del campo eléctrico que es proporcional al campo eléctrico aplicado:

$$\begin{aligned} v_{nx} &= -\mu_n E_x, \\ v_{px} &= \mu_p E_x. \end{aligned} \qquad (1.38)$$

El coeficiente de proporcional se denomina movilidad y:

- depende de la temperatura
- es inversamente proporcional a la masa efectiva del portador
- es diferente para electrones y huecos. Además:

$$\mu_n > \mu_p .$$

Conocida la velocidad de los portadores, podemos determinar fácilmente la densidad de corriente debida al arrastre de portadores. En efecto:

$$\begin{gathered} J_{nx} = qnv_x = qn\mu_n E_x \\ J_{px} = qpv_x = qp\mu_p E_x \\ \Downarrow \\ J_x = q\left(n\mu_n + p\mu_p\right)E_x \\ \Downarrow \\ J_x = \sigma E_x \,(\text{Ley de Ohm}) \end{gathered}$$

donde:

$$\sigma = q\left(n\mu_n + p\mu_p\right)$$

es la conductividad.

Para terminar este apartado dedicado a la corriente de deriva, veamos cómo se refleja en el modelo de bandas de energía la presencia de un campo eléctrico:

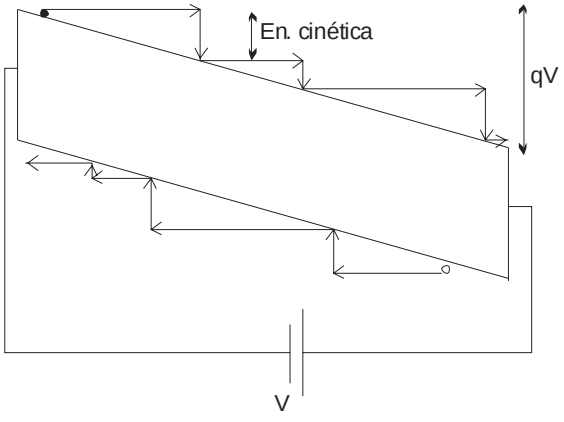


Figura 1.5.2 Inclinación del diagrama de bandas debida a la presencia de un campo eléctrico. Se representan también los procesos de aceleración y choque (pérdida de energía) que sufren los electrones

La presencia del campo eléctrico curva las bandas de energía. A la energía potencial correspondiente al mínimo de una banda hay que añadir el término debido a la energía potencial electrostática ($qV(x)$):

$$E_p(x) = E_{c0} - qV(x)\,. \tag{1.39}$$

Bajo la presencia de un campo eléctrico, los electrones son acelerados, ganando energía cinética. Después sufren dispersión y pierden su energía, siendo nuevamente acelerados.

Corriente de difusión

La corriente de difusión se debe a que una diferente concentración de portadores a lo largo del espacio provoca un flujo de éstos hacia las zonas de menor concentración y es consecuencia directa del movimiento térmico aleatorio de los portadores.

Se puede demostrar que:

$$\begin{aligned} J_{nx} &= qD_n \frac{\partial n(x)}{\partial x}, \\ J_{px} &= -qD_p \frac{\partial p(x)}{\partial x}. \end{aligned} \tag{1.40}$$

Donde:
D_n: coeficiente de difusión de los electrones

D_p : coeficiente de difusión de los huecos

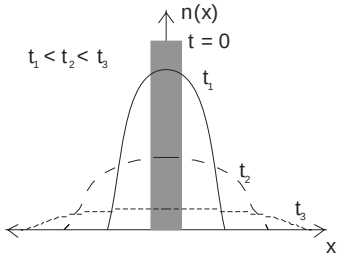


Figura 1.5.3 Evolución del exceso de portadores inicialmente presente en el origen de coordenadas al avanzar el tiempo

Corriente total. Relación de Einstein

Sumando las dos contribuciones anteriores a la corriente, la densidad de corriente de electrones y huecos viene dada por:

$$J_{nx}(x) = q\mu_n n(x)E_x(x) + qD_n \frac{\partial n(x)}{\partial x},$$
$$J_{px}(x) = q\mu_p p(x)E_x(x) - qD_p \frac{\partial p(x)}{\partial x}. \qquad (1.41)$$

Y la densidad de corriente total es:

$$J_x(x) = J_{nx}(x) + J_{px}(x). \qquad (1.42)$$

Es importante fijarse en el hecho de que la corriente arrastre es proporcional a la concentración de portadores. Por tanto, la contribución de los minoritarios a la corriente total es muy pequeña. Sin embargo, la corriente difusión es proporcional al gradiente de la concentración, por lo que la contribución de los minoritarios a la corriente puede ser considerable si hay un importante gradiente de concentración.

Relación de Einstein

En un semiconductor en equilibrio, las corrientes de difusión y arrastre (para cada tipo de portador) se deben igualar. En el caso de los electrones esto implica que:

$$q\mu_n n(x)E_x(x) = -qD_n \frac{\partial n(x)}{\partial x}. \qquad (1.43)$$

Operando se deduce de esta igualdad la Relación de Einstein entre la movilidad y el coeficiente de difusión:

$$\frac{D_n}{\mu_n} = \frac{kT}{q} \tag{1.44}$$

Análogamente, para los huecos:

$$\frac{D_p}{\mu_p} = \frac{kT}{q} \tag{1.45}$$

De la relación de Einstein se deduce que un semiconductor con buena movilidad tendrá también un alto coeficiente de difusión.

1.6 Corriente y generación-recombinación. Ecuación de continuidad

La ecuación de continuidad no es más que un balance de la variación del número de portadores con el tiempo en un elemento diferencial de volumen. Con un análisis unidimensional se obtienen las siguientes ecuaciones de continuidad para los electrones y para los huecos, respectivamente:

$$\begin{aligned} \frac{\partial n(x)}{\partial t} &= \frac{1}{q}\frac{\partial J_n(x)}{\partial x} + G - \frac{\delta n(x)}{\tau_n} \\ \frac{\partial p(x)}{\partial t} &= -\frac{1}{q}\frac{\partial J_p(x)}{\partial x} + G - \frac{\delta p(x)}{\tau_p} \end{aligned} \tag{1.46}$$

Es decir, la variación en la concentración de portadores por unidad de tiempo tiene varias contribuciones: la debida a la variación de la corriente, más la de los portadores que se generan menos los que se recombinan (se ha supuesto la aproximación de baja inyección).

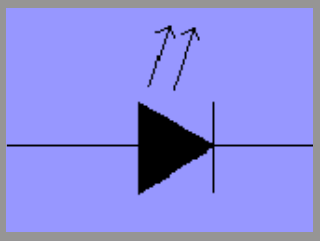
REFERENCIAS

[1] B.G. Streetman, S. Banerjee, “Solid State Electronic Devices”, Prentice Hall, 2000.
[2] S. Dimitrijev, “Understanding Semiconductor Devices”, Oxford University Press, 2000.

2

Capítulo

Uniones



Símbolo de un LED

Índice

Objetivos

- Explicar cualitativamente el funcionamiento de un diodo de unión PN.
- Obtener la expresión que relaciona la corriente con la tensión aplicada en condiciones estacionarias.
- Obtener modelos de gran señal.
- Estudiar la conmutación del diodo entre los estados de conducción y no corte.
- Proporcionar un modelo de pequeña señal.
- Simplificar los modelos equivalentes encontrados para facilitar el análisis de estos circuitos a mano.
- Estudiar otras uniones: unión metal-semiconductor y heterouniones

Palabras Clave

Diodo.
Unión PN.
Unión metal-semiconductor.
Heterounión.
Rectificación.
Zéner.
LED.
Varactor.
Fotodiodo.
Modelo de gran señal.
Modelo de pequeña señal.
Capacidades de las uniones.
Resistencias parásitas.
Respuesta en frecuencia.
Polarización.
Circuitos con diodos.

2.1 Introducción

Los diodos son componentes electrónicos no lineales que, idealmente, dejan pasar la corriente en un solo sentido. Usualmente están basados en la unión de dos semiconductores, uno tipo N y otro tipo P. Después se verán otros tipos de uniones y diodos.

La unión PN también aparecerá en el estudio de otros dispositivos y en la fabricación de circuitos integrados, por lo que resulta interesante estudiarla no sólo por sus aplicaciones como diodo.

El símbolo de un diodo es:

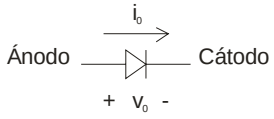


Figura 2.1.1 Símbolo de un diodo

La característica de transferencia de un diodo ideal se representa en la Figura 2.1.2.

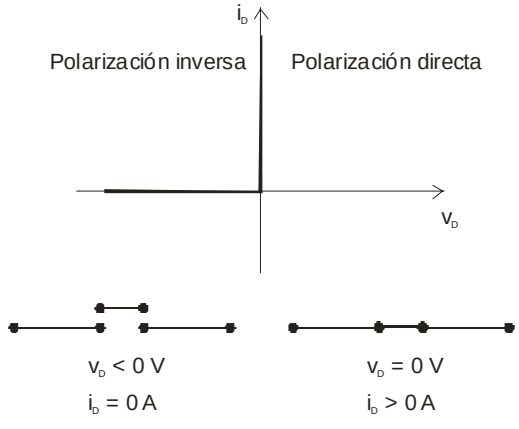


Figura 2.1.2 Característica de transferencia de un diodo ideal. Cuando la tensión aplicada entre sus extremos es negativa, se comporta como un circuito abierto impidiendo el paso de corriente (independientemente de la tensión aplicada). Sin embargo, permite la circulación de corriente en sentido contrario sin caída de tensión (se comporta como un cortocircuito).

Entre las principales aplicaciones de los diodos cabe destacar su uso:

- En rectificadores
- Como diodos emisores de luz (LEDs)

- Como detectores de luz
- Como capacidades dependientes de tensión (varactores)
- En osciladores (diodos túnel)

2.2 Unión PN en equilibrio térmico

Descripción

En condiciones de equilibrio térmico no hay excitación externa sobre el dispositivo y no debe haber ninguna corriente neta.

Para analizar qué ocurre en una unión PN en equilibrio térmico supongamos primero las dos zonas P y N por separado:

- o zona N: hay una gran concentración de electrones
- o zona P: hay una gran concentración de huecos

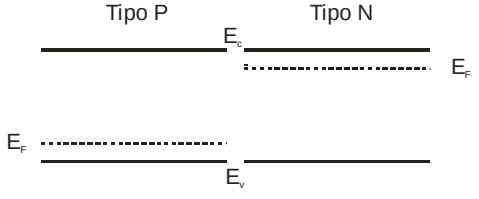


Figura 2.2.1 Diagramas de bandas de un semiconductor tipo P y de un tipo N por separado.

Al poner los dos semiconductores en contacto, tendremos corrientes de difusión de electrones y huecos próximos a la unión que tienden a igualar las concentraciones. Este proceso no puede seguir indefinidamente porque se generaría carga no balanceada en las dos regiones.

El campo generado por las cargas fijas que dejan los portadores al difundirse se opone a este movimiento, generando una barrera de potencial que se opone a la difusión. En sentido contrario a la corriente de difusión, tenemos una de arrastre provocada por este campo eléctrico. Ambas se compensan en equilibrio térmico, de forma que no hay trasporte neto de carga (corriente eléctrica nula).

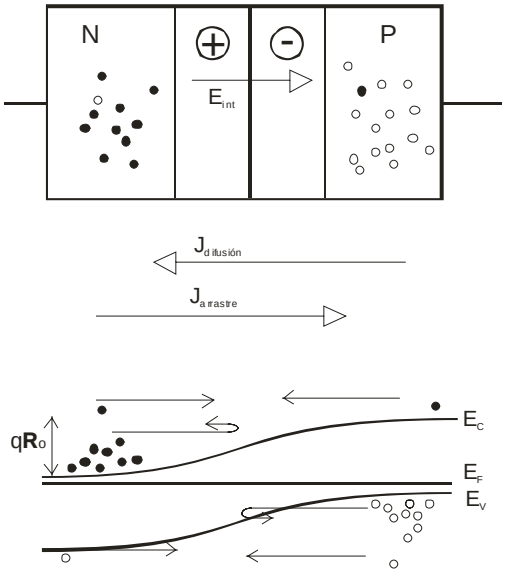


Figura 2.2.2 El campo eléctrico que aparece en la unión se opone a la difusión de los portadores mayoritarios

Los electrones (y huecos) que pueden superar la barrera de potencial y difundirse se compensan exactamente con los pocos que hay en la zona P (o zona N para los huecos) y que son arrastrados por el campo eléctrico. Por tanto, como hemos dicho, en equilibrio no hay corriente neta y se verifican las siguientes igualdades:

$$\begin{aligned} J_n(deriva) + J_n(difusion) &= 0 \\ J_p(deriva) + J_p(difusion) &= 0. \end{aligned} \qquad (2.1)$$

Se denomina zona de carga espacial a la zona en la que hay campo eléctrico e impurezas donadoras o aceptadoras sin cancelar. La concentración de portadores móviles en esta zona es muy pequeña puesto que son barridos por el campo eléctrico.

Cálculo del potencial de unión

Definimos los potenciales de Fermi[1] como la distancia desde el nivel de Fermi hasta el nivel de Fermi intrínseco (como potenciales). Por tanto, estos potenciales de Fermi verifican:

$$\begin{aligned} q\phi_{FN} &= E_{FN} - E_{iN} \qquad \text{(zona N)}, \\ q\phi_{FP} &= -\left(E_{FP} - E_{iP}\right) \qquad \text{(zona P)}. \end{aligned} \qquad (2.2)$$

[1]De aquí en adelante, el subíndice *N* se usará para indicar que la magnitud a la que acompaña se refiere a la zona N de la unión. Análogamente, el subíndice *P* indica que la variable en cuestión se refiere a la zona P.

A partir de estas definiciones, la altura de la barrera de energía potencial (qV_0) viene dada por (ver Figura 2.2.3):

$$qV_0 = q\phi_{FN} + q\phi_{FP} \,. \tag{2.3}$$

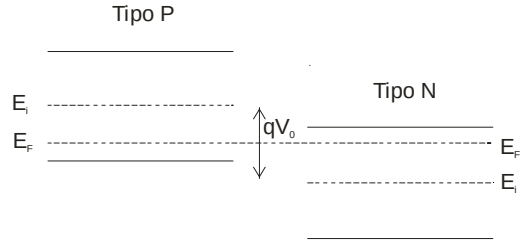


Figura 2.2.3 Cálculo del potencial barrera (o de unión) como suma de los potenciales de Fermi

Ya conocemos la altura de la barrera de potencial en función de los pseudoniveles de Fermi. Pero resulta más interesante calcularla en función de los dopados de los dos semiconductores. Para ello, sólo queda relacionar los potenciales de Fermi con las concentraciones de portadores o con los dopados:

$$\begin{aligned} n_{N0} &= n_i e^{\frac{q\phi_N}{kT}}, \\ p_{P0} &= n_i e^{\frac{q\phi_P}{kT}}. \end{aligned} \tag{2.4}$$

Luego:

$$\mathrm{V}_0 = \frac{kT}{q} \ln\left(\frac{n_{N0} p_{P0}}{n_i^2}\right) \approx \frac{kT}{q} \ln\left(\frac{N_D N_A}{n_i^2}\right). \tag{2.5}$$

Resulta interesante relacionar los cocientes de las concentraciones de electrones (o huecos) en las dos zonas neutras (N y P). De la primera igualdad de (2.5) y aplicando la ley de acción de masas:

$$\frac{n_{N0}}{n_{P0}} = \frac{P_{P0}}{p_{N0}} = e^{qV_0/kT} \,. \tag{2.6}$$

Ejercicio

Demostrar las igualdades (2.5) y (2.6) imponiendo que en equilibrio la corriente es cero (se cancelan las corrientes de difusión y deriva, tanto para electrones como para huecos).

2.3 Unión PN polarizada en condiciones estacionarias

Descripción cualitativa

En equilibrio térmico hemos visto que la altura de la barrera es tal que se compensan las corrientes de difusión y deriva.

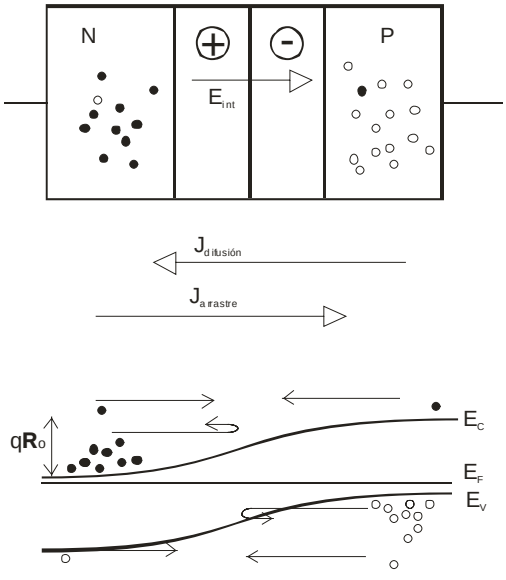


Figura 2.3.1 Unión PN en equilibrio térmico: las corrientes de difusión y deriva se cancelan

Si se aplica una tensión externa V_D, se reduce (si $V_D > 0$) o se incrementa la barrera ($V_D < 0$). Para el análisis de la unión PN polarizada, supondremos que toda la tensión aplicada V_D cae en la región de la unión y una cantidad despreciable en las zonas neutras.

Polarización inversa

Cuando se aplica una tensión inversa (negativa) se incrementa la barrera de potencial respecto al caso de equilibrio térmico y, como consecuencia, se dificulta la corriente de difusión. Sin embargo, la corriente de arrastre apenas aumenta porque no está limitada por la velocidad de los portadores, sino por su cantidad. En consecuencia: *la magnitud de V_D en inversa apenas influye sobre la corriente*. Es decir, con tensiones inversas se obtiene una corriente $-I_S$ independiente de la polarización.

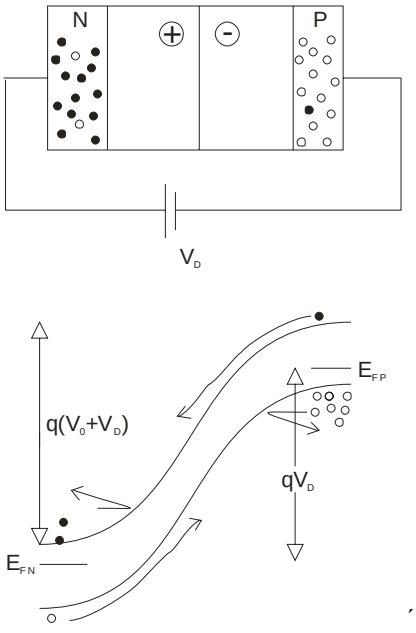


Figura 2.3.2 Polarización inversa: la barrera de potencial aumenta. Se bloquea la difusión de mayoritarios y no se altera la corriente de arrastre, que está limitada por la generación de minoritarios.

Polarización directa

Con tensiones positivas, disminuye la barrera que bloquea la difusión de los portadores mayoritarios. Por tanto:

$$\begin{aligned} J_n(difusion) > J_n(deriva), \\ J_p(difusion) > J_p(deriva). \end{aligned} \qquad (2.7)$$

Cuanto más se disminuya la barrera, más portadores mayoritarios la podrán superar y difundirse hacia el otro lado de la unión y por tanto mayor será la corriente.

Los electrones y huecos inyectados (desde zona N y P, respectivamente, donde son mayoritarios) se convierten en portadores minoritarios al atravesar la unión, incrementando notablemente la concentración de minoritarios en la zona P y N, respectivamente.

La concentración de mayoritarios apenas cambia, puesto que aumenta en la misma cantidad que la de minoritarios (para mantener la neutralidad eléctrica) y trabajaremos bajo la hipótesis de baja inyección (ver Figura 2.3.4).

La corriente en las zonas neutras se debe a la difusión de los portadores minoritarios y a los mayoritarios que van a recombinarse con ellos.

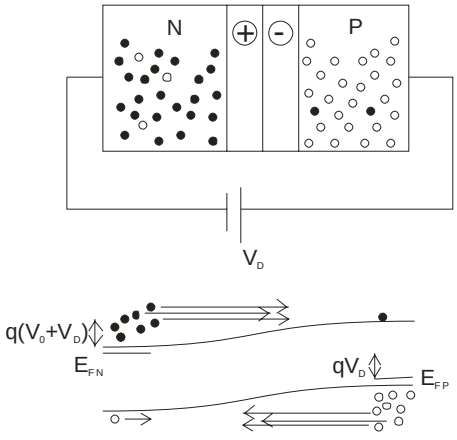


Figura 2.3.3 Polarización directa: disminución de la barrera de potencial y aumento de la corriente de difusión.

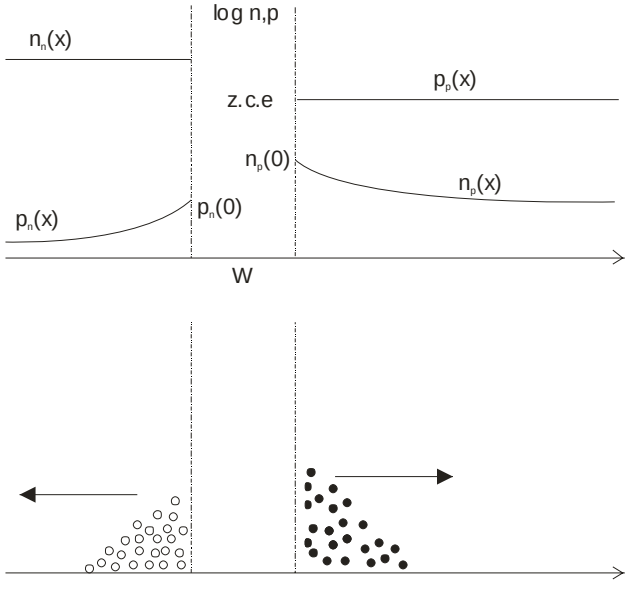


Figura 2.3.4 Los portadores inyectados se convierten en minoritarios e incrementan la concentración de éstos. La de mayoritarios apenas se altera (a). La corriente en el diodo se debe a la difusión y recombinación de estos portadores minoritarios (b).

Curva I-V estática

Como demostraremos en el siguiente apartado, la dependencia entre la corriente que circula por el diodo y la tensión aplicada entre sus extremos viene dada por:

$$I_D = I_S\left[e^{\frac{V_D}{V_T}} - 1\right], \tag{2.8}$$

con:

$$\begin{aligned} I_S &= qA\left(\frac{D_p}{L_p}p_{n0} + \frac{D_n}{L_n}n_{p0}\right), \\ V_T &= \frac{kT}{q}, \\ A &= \text{área de la unión.} \end{aligned} \tag{2.9}$$

En la Figura 2.3.5 se muestra la representación gráfica de esta curva I-V (ecuación (2.8)). Como puede verse, cuando la tensión es negativa, la corriente es muy pequeña e independiente de la tensión. Sin embargo, para valores positivos la corriente aumenta notablemente con pequeños incrementos de tensión.

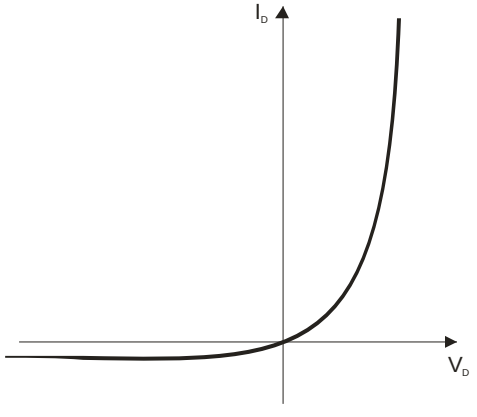


Figura 2.3.5 Característica I-V de una unión PN

Descripción cuantitativa. Cálculo de la curva I-V

En este apartado vamos a demostrar la expresión (2.8). La corriente se debe a la difusión y recombinación de los portadores minoritarios inyectados desde el otro lado de la unión. Fijémonos, por ejemplo, en los huecos inyectados en la zona N. La corriente de difusión en un cierto punto *x* viene dada por:

$$j_p = -qD_p \frac{\partial p_n(x)}{\partial x}. \tag{2.10}$$

En el capítulo anterior se demostró que la distribución $p_n(x)$ cuando se inyectan portadores que se difunden y recombinan viene dada por el siguiente exceso de portadores:

$$\delta p_n(x) = \Delta p_n e^{\frac{-(x_n - x_{n0})}{L_p}}. \tag{2.11}$$

Evaluando la corriente justo en x_{n0} se obtiene, usando (2.10) y (2.11), la siguiente expresión:

$$j_p = qD_p \frac{\Delta p_p}{L_p}. \tag{2.12}$$

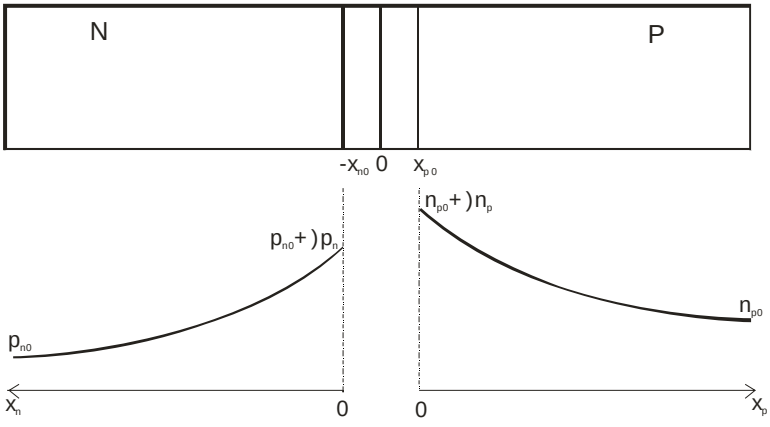


Figura 2.3.6 Esquema del exceso de minoritarios en función de la posición.

Para conocer la corriente, ya sólo nos falta calcular Δp_n. Si en equilibrio térmico:

$$\frac{p_{p0}}{p_{n0}} = e^{V_0/V_T}, \tag{2.13}$$

con polarización se debe cumplir que:

$$\frac{p_p(-x_{p0})}{p_n(x_{n0})} = e^{(V_0 - V_D)/V_T}. \tag{2.14}$$

Desarrollando:

$$\frac{p_p(-x_{p0})}{p_n(x_{n0})} = \frac{p_{p0}}{p_{n0}} e^{-V_D/V_T}. \tag{2.15}$$

La concentración de mayoritarios apenas se ve alterada, por lo que:

$$p_p(-x_{p0}) \approx p_{p0}.$$

y por tanto:

$$\frac{p_n(x_{n0})}{p_{n0}} \approx e^{V_D/V_T} . \qquad (2.16)$$

(Como ya habíamos discutido de forma cualitativa, vemos ahora de nuevo que la población de minoritarios aumenta o disminuye respecto del equilibrio térmico dependiendo de la tensión V_D). Finalmente:

$$\Delta p_n = p_n(x_{n0}) - p_{n0} = p_{n0}(e^{V_D/V_T} - 1) . \qquad (2.17)$$

Sustituyendo (2.17) en (2.12) obtenemos la corriente de difusión de los minoritarios en la zona N:

$$j_p = q\frac{D_p}{L_p} p_{n0}(e^{V_D/V_T} - 1) . \qquad (2.18)$$

Análogamente, la difusión de electrones en la zona P es:

$$j_n = q\frac{D_n}{L_n} n_{p0}(e^{V_D/V_T} - 1) . \qquad (2.19)$$

Por tanto, quedan demostradas las expresiones (2.8) y (2.9).

Observaciones

- Si el dopado de las dos zonas es diferente, la corriente está determinada principalmente por la zona más dopada (es la que inyecta más minoritarios).
- Calculada la corriente de minoritarios (difusión) la de mayoritarios es fácil de obtener, puesto que la corriente *I* debe ser constante a lo largo de todo el diodo (ver Figura 2.3.7)
- En la demostración de la curva I-V (ecuaciones (2.8) y (2.9)) hemos asumido implícitamente las siguientes aproximaciones:
  - Toda la tensión cae en la unión (resistencia despreciable en las zonas neutras). Esto es cierto si:
    - El nivel de inyección de portadores es bajo.
    - La resistencia de las zonas neutras es baja:
  - No hay generación-recombinación en la zona de transición

Hay que destacar que a pesar de la primera suposición, los mayoritarios se mueven mediante arrastre por el campo eléctrico en las zonas neutras (pero, como hay muchos mayoritarios, el campo eléctrico necesario para generar una corriente apreciable puede considerarse despreciable).

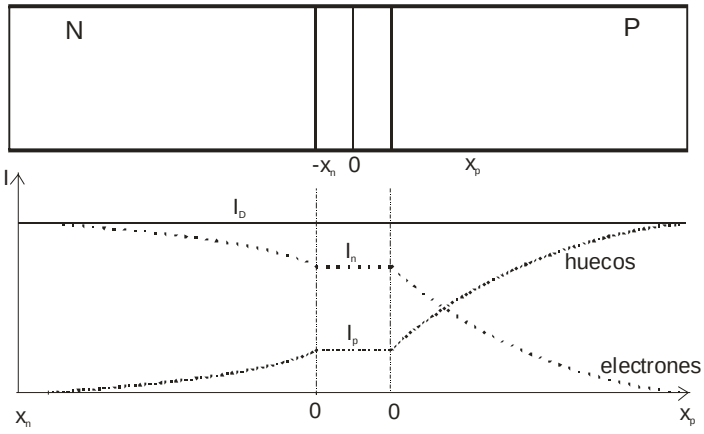


Figura 2.3.7 La corriente en el dispositivo es la misma en todas las posiciones. La corriente de arrastre puede calcularse como la corriente total menos la corriente de difusión en el punto considerado. La corriente total viene dada por la suma de las corrientes de difusión justo en los extremos de las zonas neutras con la zona de carga espacial.

- **Polarización inversa.** Las expresiones (2.11) y (2.17) siguen siendo ciertas pero con V_D negativo:

$$\begin{aligned} \delta p_n(x) &= \Delta p_n e^{\frac{-(x_n - x_{n0})}{L_p}}, \\ \Delta p_n &= p_{n0}(e^{-|V_D|/V_T} - 1) \approx -p_{n0}. \end{aligned} \tag{2.20}$$

Los minoritarios próximos a la zona de carga espacial son barridos por el campo eléctrico hacia el otro lado de la barrera. Esta corriente está limitada por lo rápido que se repongan (mediante generación térmica) los minoritarios barridos[2].

---

[2]Obsérvese que la tasa de generación por unidad de volumen es $g = \frac{p_{n0}}{\tau_p}$.

Si asumimos que todos los huecos generados en el volumen AL_P son inyectados hacia el otro lado de la unión tenemos una corriente igual a:

$$\left|I_p\right| = qAL_p \frac{p_{n0}}{\tau_p},$$

que coincide con la expresión de I_S. Vemos de este modo cuál es la interpretación de la corriente inversa de saturación y cómo está limitada por la generación térmica de portadores (recordar que $L_p^2 = D_p \tau_p$).

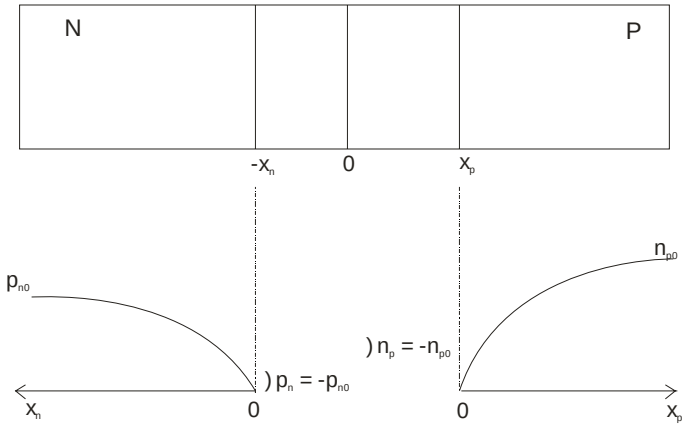


Figura 2.3.8 Concentración de minoritarios con polarización inversa. Un defecto de minoritarios (que son barridos hacia el otro lado de la barrera por el campo eléctrico) viene acompañado del mismo defecto de portadores en la población de mayoritarios.

2.4 Dependencia con la temperatura

Aparte de la dependencia explícita de la corriente que circula por el diodo con la temperatura (que se manifiesta en las expresiones (2.8) y (2.9)), es más fuerte la dependencia de I_S con la temperatura. En efecto, la corriente inversa de saturación es directamente proporcional a:

$$I_S \propto n_i^2 = N_c N_v e^{-E_g / kT}. \qquad (2.21)$$

La influencia total de la temperatura sobre la corriente que circula por el diodo básicamente consiste en un desplazamiento de la curva I-V en el eje V hacia la izquierda (unos - 2 mV/ºC), como se ilustra en la Figura 2.5.1.

2.5 Modelos I-V de gran señal. Análisis de circuitos con diodos

En el apartado anterior ya hemos comprobado que la relación entre la intensidad que circula por el diodo y la tensión que se aplica es de tipo exponencial y viene dada por:

$$I = I_S\left(e^{V/V_T} - 1\right). \qquad (2.22)$$

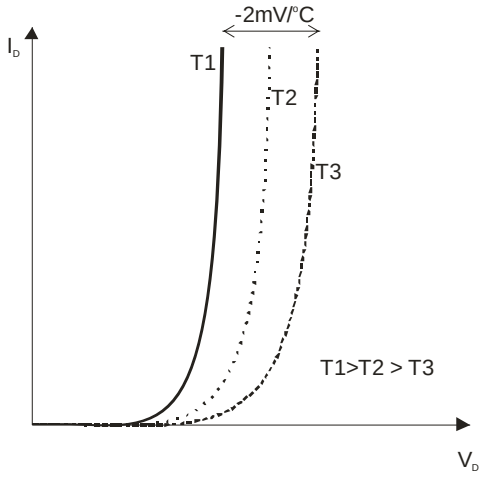


Figura 2.5.1 Dependencia de la curva I-V con la temperatura.

Ejercicio

Calcular la intensidad que circula por el circuito de la Figura 2.5.2 y la tensión que cae en el diodo. Datos: I_S = 2 fA, V_{DD} = 5 V y R = 1KΩ.

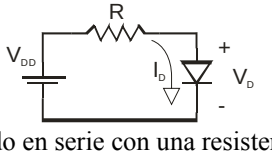


Figura 2.5.2 Diodo en serie con una resistencia. Sólo en circuitos sencillos puede usarse la ecuación .(2.22) Para la resolución a mano del circuito.

Como se habrá comprobado al resolver el ejercicio, la ecuación del diodo no es cómoda para resolver circuitos a mano, aunque sí se usa para la resolución de circuitos mediante ordenador (p.e., con Spice).

Para cálculos a mano se usan los modelos simplificados que se muestran en la Figura 2.5.3.

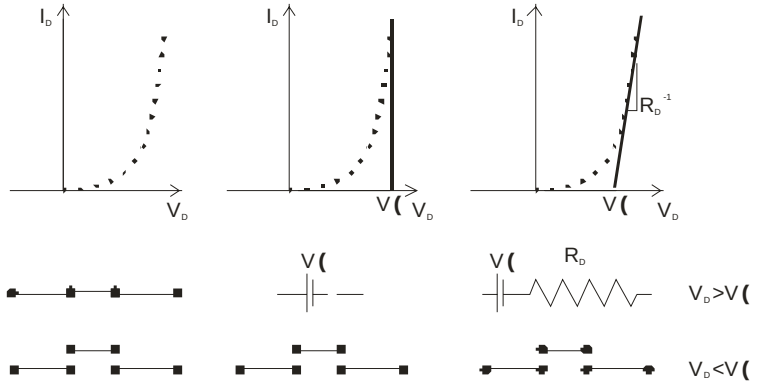


Figura 2.5.3 Modelos simplificados de la curva I-V de un diodo.

Ejercicio

Resolver el ejercicio anterior empleando cada uno de los modelos de la Figura 2.5.3. Comparar los resultados. Datos: $V_\gamma = 0.65$ V, $R_D = 20\ \Omega$.

2.6 Distribución de carga y campo en la unión. Cálculo del ancho de la zona de carga espacial

Para el cálculo de la carga que hay en la unión, supondremos las siguientes aproximaciones:

- Unión abrupta y dopados uniformes
- Toda la carga de la zona de carga espacial es fija (se debe a las impurezas ionizadas), no hay carga debida a electrones o huecos (aproximación de deplexión o vaciamiento)
- Todas las impurezas están ionizadas

Hay el mismo número de cargas negativas (impurezas aceptadoras ionizadas) que de cargas positivas (impurezas donadoras ionizadas) en la zona de carga espacial (en caso contrario, el campo eléctrico no sería nulo en las zonas neutras). Por tanto:

$$x_{n0} N_D = x_{p0} N_A . \tag{2.23}$$

Esto implica, por ejemplo, que:

$$N_A >> N_D \Rightarrow x_{n0} >> x_{p0} . \tag{2.24}$$

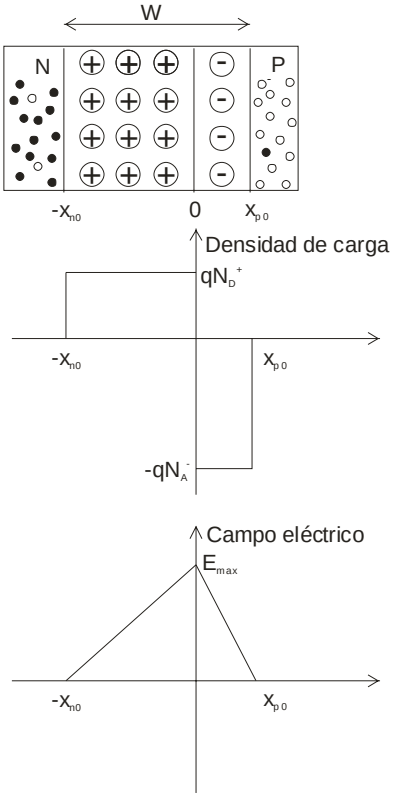


Figura 2.6.1 a) Zona de carga espacial en una unión PN abrupta. b) Densidad de carga en la z.c.e. c) Campo eléctrico.

Cálculo del campo eléctrico

El campo eléctrico lo obtendremos integrando una vez la ecuación de Poisson unidimensional a lo largo de la zona de carga espacial:

$$\frac{d^2V(x)}{d^2x} = \frac{-\rho(x)}{\varepsilon}. \qquad (2.25)$$

($\rho(x)$ es la densidad de carga y ε la constante dieléctrica del semiconductor). Se obtiene:

$$\begin{aligned} \frac{dE(x)}{dx} &= \frac{-q}{\varepsilon} N_A \quad (\text{si } 0 > x > x_{p0}) \\ \frac{dE(x)}{dx} &= \frac{q}{\varepsilon} N_D \quad (\text{si} - x_{n0} < x < 0) \end{aligned} \qquad (2.26)$$

Integrando las anteriores ecuaciones entre un borde de la zona de carga espacial y el punto de unión (o por geometría, observando la Figura

2.6.1c) se obtiene el campo eléctrico máximo:

$$E(x=0) = E_{\max} = \frac{qN_D x_{n0}}{\varepsilon} = \frac{qN_A x_{p0}}{\varepsilon}. \qquad (2.27)$$

Ancho de la z.c.e:

A continuación, se integra la expresión:

$$\frac{dV(x)}{dx} = -E(x), \qquad (2.28)$$

para conseguir eliminar la dependencia de $E(x)$ con la posición y relacionar E y la anchura de la z.c.e. con el potencial barrera:

$$\int_{-x_{p0}}^{x_{n0}} \frac{dV(x)}{dx} dx = \int_{-x_{p0}}^{x_{n0}} -E(x)dx$$

$$\Downarrow$$

$$V(x_{n0}) - V(-x_{p0}) = -\int_{-x_{p0}}^{x_{n0}} E(x)dx. \qquad (2.29)$$

$$\Downarrow$$

$$V_0 = -\int_{-x_{p0}}^{x_{n0}} E(x)dx$$

La integral del término de la derecha es igual al área que encierra la gráfica de $E(x)$ (Figura 2.6.1c). Por tanto:

$$V_0 = \frac{1}{2}(x_{n0} + x_{p0})|E_{\max}|. \qquad (2.30)$$

Si definimos $W = x_{n0} + x_{p0}$, entonces:

$$V_0 = \frac{1}{2} W E_{\max} = \frac{1}{2} \frac{qN_D x_{n0}}{\varepsilon} W. \qquad (2.31)$$

Teniendo en cuenta (2.23) y operando obtenemos:

$$V_0 = \frac{1}{2} \frac{q}{\varepsilon} \frac{N_A N_D}{N_A + N_D} W^2. \qquad (2.32)$$

Por tanto:

$$W = \sqrt{\frac{2\varepsilon V_0}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)}$$

$$x_{p0} = \frac{WN_D}{N_A + N_D}. \qquad (2.33)$$

$$x_{n0} = \frac{WN_A}{N_A + N_D}$$

Observaciones

- Si en lugar de estar en condiciones de equilibrio térmico se aplica un voltaje V_D, entonces $V(x_{n0}) - V(-x_{p0}) = V_0 - V_D$ y se obtiene:

$$W = \sqrt{\frac{2\varepsilon(V_0 - V_D)}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)}\ . \qquad (2.34)$$

- W crece al aumentar el voltaje aplicado en inversa (V_D negativo).
- Si un lado de la unión está mucho más dopado que el otro (por ejemplo, $N_A >> N_D$) entonces:
  - W depende de la concentración de las impurezas del lado menos dopado. En el caso supuesto:

$$W \approx \sqrt{\frac{2\varepsilon(V_0 - V_D)}{q}\frac{1}{N_D}}.$$

  - Casi toda la z.c.e. está en la región del semiconductor menos dopado (como también se deduce de (2.23)):

$$x_{p0} = \frac{WN_D}{N_A + N_D} \approx 0,$$

$$x_{n0} = \frac{WN_A}{N_A + N_D} \approx W.$$

2.7 Fenómenos de ruptura

Hasta ahora hemos visto que los diodos en inversa sólo conducen con una corriente muy pequeña ($-I_S$), denominada corriente inversa de saturación, que es independiente de la tensión aplicada. Sin embargo, realmente no se puede aplicar cualquier voltaje en inversa manteniéndose esta situación, sino que a partir de cierto valor crítico o tensión de ruptura, la corriente inversa del diodo se incrementa abruptamente, obteniéndose un rango muy elevado de corriente con una pequeña variación de la tensión (de forma análoga a como sucede en la región de conducción directa). La Figura 2.8.1 muestra las dos regiones de conducción de una unión PN.

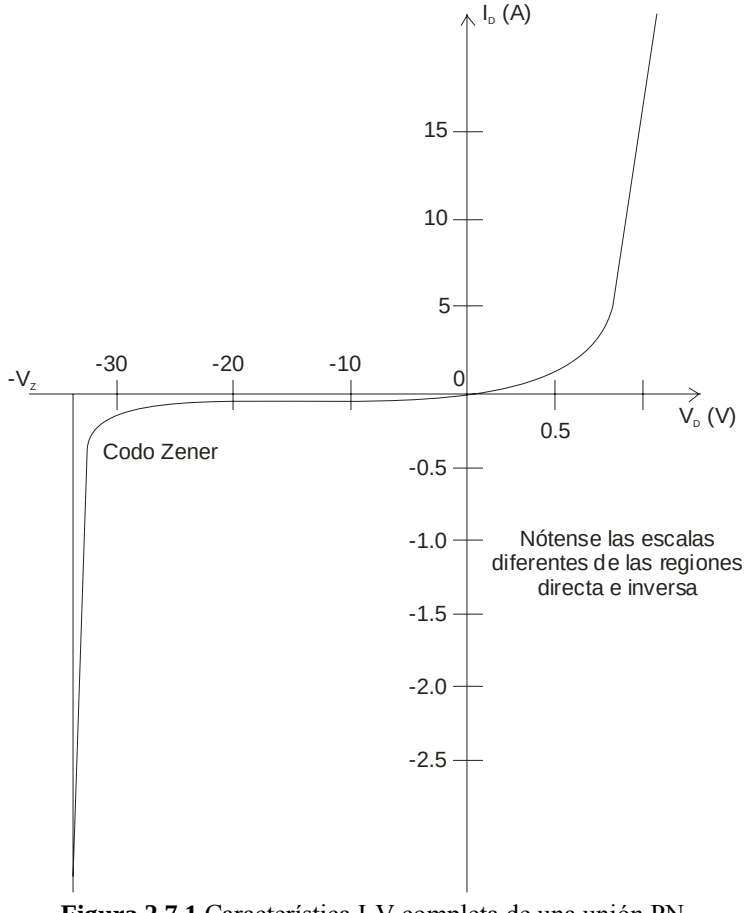


Figura 2.7.1 Característica I-V completa de una unión PN.

A pesar del nombre de ruptura, este comportamiento no es destructivo en sí mismo si se limita la corriente (con una resistencia externa, por ejemplo), al igual que sucede en la región directa de conducción.

Los diodos especialmente diseñados para conducir en inversa se suelen emplear para regular el voltaje en un punto (por ejemplo, en fuentes de tensión) y se representan con un símbolo ligeramente diferente al de un diodo (Figura 2.7.2).

Figura 2.7.2 Símbolo de un diodo Zener

La ruptura o conducción en inversa puede deberse a dos fenómenos independientes: la ruptura por avalancha y la ruptura Zéner o túnel.

La ruptura por avalancha se ilustra en la Figura 2.7.3. Ya hemos visto que en inversa la corriente se debe a los electrones y huecos minoritarios que son arrastrados por el campo eléctrico de la zona de carga espacial. Sigamos el camino de un electrón. La energía cinética de un electrón se incrementa conforme avanza a causa de la aceleración provocada por este campo eléctrico. Puede ser que mientras se mueve, el electrón choque con los átomos de la red cristalina. Si ha adquirido suficiente energía cinética podría incluso romper un enlace entre átomos de silicio, generando un par electrón-hueco. Los dos electrones (el inicial y el generado en el choque) y el hueco son acelerados de nuevo por el campo eléctrico, incrementando la corriente inversa. Además, los dos electrones pueden adquirir de nuevo la suficiente energía como para generar cada uno otro par electrón-hueco y así sucesivamente en un proceso en cadena que puede llegar a incrementar notablemente la corriente inversa.

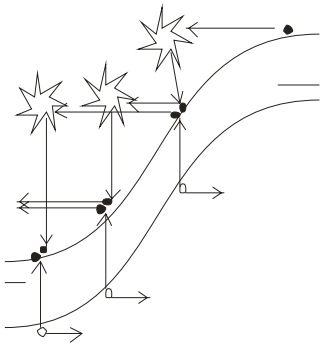


Figura 2.7.3 Ruptura por avalancha en una unión PN.

La corriente por avalancha tiene un coeficiente de temperatura negativo. Es decir, la corriente disminuye al aumentar la temperatura. Esto se debe a que cuanto mayor sea la temperatura, mayor es la agitación térmica de los átomos y, por tanto, mayor es la probabilidad de choque de éstos con los electrones. Esto hace que los electrones tengan menos opciones de adquirir la energía suficiente para ser capaces de romper los enlaces y generar nuevos pares electrón-hueco.

El otro proceso de ruptura, la rotura Zéner o túnel, se ilustra en la Figura 2.7.4. La rotura túnel tiene lugar si la distancia entre las bandas de conducción y valencia es pequeña en la zona de carga espacial. Esto sucede cuando el dopado de los semiconductores es alto, ya que esto favorece que la anchura de la zona de carga espacial sea pequeña y la curvatura de las bandas tenga lugar en poco espacio.

Mediante el proceso túnel un electrón puede pasar de la banda de valencia a la de conducción (Figura 2.7.4). Este proceso es un mecanismo de origen cuántico y no tiene explicación dentro del contexto de la física clásica. Debido a que los electrones también tienen un carácter ondulatorio pueden atravesar, con cierta probabilidad, barreras de potencial si éstas son lo suficientemente bajas o estrechas, como es el caso.

Al contrario que la ruptura por avalancha, la ruptura Zener tiene un coeficiente de temperatura positivo, puesto que se facilita la ruptura de un enlace al aumentar la temperatura.

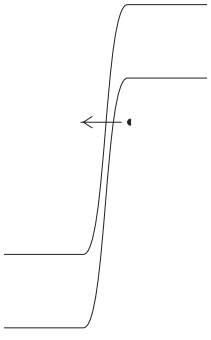


Figura 2.7.4 Ruptura mediante proceso túnel

El que tenga lugar en primer lugar la ruptura por avalancha o por efecto túnel está básicamente controlado por el dopado de los semiconductores. Se puede conseguir que la tensión de ruptura sea casi independiente de la temperatura si se diseña la unión de forma que la ruptura tenga lugar por ambos mecanismos a la vez.

Como el lector habrá observado, a los diodos que trabajan en la región inversa se les denomina diodos Zéner independientemente del fenómeno en concreto que tenga lugar (podría ser únicamente la ruptura por avalancha, como de hecho suele preferirse por presentar una característica más abrupta).

2.8 Comportamiento dinámico. Modelos de conmutación y de pequeña señal

Introducción

Hasta ahora hemos estudiado la unión PN en condiciones estacionarias. Sin embargo, normalmente los diodos se usan en aplicaciones de conmutación o en procesado de señales variables, por lo que se requiere un análisis del comportamiento transitorio y de pequeña señal.

Variación temporal de la carga móvil almacenada

El comportamiento temporal viene determinado por el tiempo necesario para eliminar o crear el exceso de carga en las zonas neutras. Para determinarlo, se necesitan ecuaciones dependientes del tiempo y, por tanto, se recurrirá a la ecuación de continuidad.

Antes de ello, debemos tener en mente que la corriente I_D en estado estacionario depende del exceso de portadores minoritarios en las zonas neutras y, por tanto, cualquier variación de I_D (por ejemplo, en un transitorio de corte) implica la variación del exceso de carga. Por otro lado, aumentar o disminuir el exceso de carga requiere tiempo y por ello el cambio de estado de un diodo no es un proceso instantáneo.

Como hemos dicho, para analizar la evolución temporal se usa la ecuación de continuidad. Una vez integrada en el espacio (esto es, en la variable x_n) se obtiene (para los huecos):

$$i_p(t) = \frac{Q_p(t)}{\tau_p} + \frac{dQ_p(t)}{dt}. \qquad (2.35)$$

donde:

- $i_p(t)$ es la corriente debida a los huecos
- $Q_p(t)$ es la carga total en exceso o defecto (debida a los huecos) almacenada en la zona neutra N
- τ_p es el tiempo de vida media (o de recombinación) de los huecos en la zona N

La ecuación (2.35) muestra que la corriente tiene dos contribuciones: aportar los huecos que se recombinan e incrementar la carga almacenada.

Análogamente, se obtiene otra expresión para los electrones. Por simplicidad, en lo sucesivo supondremos una unión P^+N (esto es, $N_A >> N_D$), de forma que la corriente se debe casi exclusivamente a la contribución de los huecos ($i(t) \approx i_p(t)$).

La ecuación (2.35) determina el comportamiento transitorio del diodo. Su solución depende del tipo concreto de transición que tenga lugar. Veremos dos:

- Transitorio de corte (paso de polarización directa a corte)
- Transitorio de paso a inversa

Conmutación. Transitorio de corte

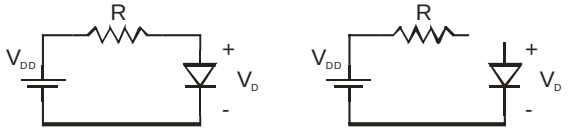


Figura 2.8.1 El transitorio de corte del diodo transcurre entre las situaciones estacionarias representadas en las dos figuras.

En este transitorio, en el instante $t = 0$, se abre el circuito y se impide el paso de corriente por el diodo.

En la situación estacionaria, antes del corte, se verifica que:

$$i(t<0) = I_i = \frac{Q_p(0)}{\tau_p}, \qquad (2.36)$$

ya que $\frac{dQ_p(t)}{dt} = 0$). Despejando obtenemos la siguiente condición inicial que usaremos posteriormente:

$$Q_p(0) = I_i \tau_p . \qquad (2.37)$$

Para instantes $t > 0$, la ecuación diferencial (2.35) cumple:

$$i(t>0) = 0 = \frac{Q_p(t)}{\tau_p} + \frac{dQ_p(t)}{dt} \Rightarrow \frac{dQ_p(t)}{dt} = -\frac{Q_p(t)}{\tau_p} . \qquad (2.38)$$

Resolviendo esta ecuación diferencial (con la condición inicial dada por (2.37)) obtenemos el exceso de carga en función del tiempo:

$$\begin{aligned} Q_p(0) &= I_i \tau_p \\ Q_p(t>0) &= Q_p(0) e^{-t/\tau_p} . \end{aligned} \qquad (2.39)$$

Como vemos, el exceso de portadores minoritarios desaparece conforme avanza el tiempo a una velocidad controlada por τ_p (como es lógico, puesto que el exceso de carga desaparece por recombinación, no circula corriente).

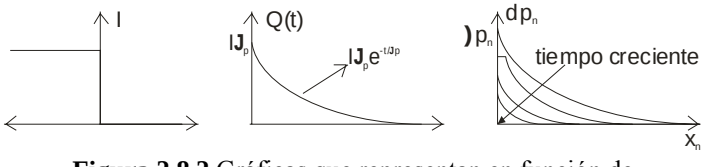


Figura 2.8.2 Gráficas que representan en función de tiempo, en un transitorio de corte, la intensidad (a), la carga total almacenada (b) y la evolución del exceso de portadores (c).

Por otro lado, ya se ha visto que el exceso de portadores está relacionado con la caída de potencial en el diodo mediante la siguiente expresión:

$$\Delta p_n(t) = p_{n0}(e^{v_D(t)/V_T} - 1). \qquad (2.40)$$

Por tanto, aunque desde $t = 0$ no circule corriente, la tensión v_D no se anula hasta que desaparece el exceso de minoritarios (huecos en nuestro caso particular).

En la figura se muestra cómo evoluciona la tensión v_D para diferentes valores del tiempo de recombinación, que es el que determina la rapidez con la que desaparece el exceso de minoritarios.

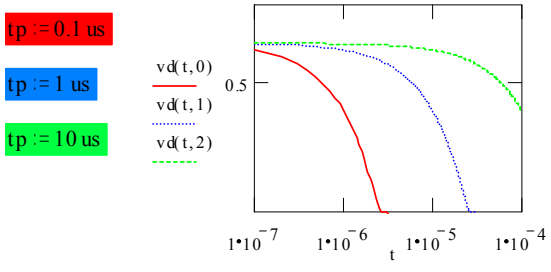


Figura 2.8.3 Gráfica de Mathcad que muestra la evolución temporal de la tensión V_D para tres valores diferentes de τ_p.

Como puede verse, para mejorar la rapidez del diodo hay que disminuir τ_p. Esto puede lograrse introduciendo impurezas que favorecen la recombinación (como el oro) al crear niveles energéticos aproximadamente en la mitad de la banda prohibida. Otra opción es usar diodos cortos, que acumulan poca carga en las regiones neutras.

Conmutación. Transitorio de paso a inversa

Es el que tiene lugar en la mayoría de las aplicaciones de conmutación. Para su estudio, supongamos el circuito mostrado en la Figura 2.8.4.

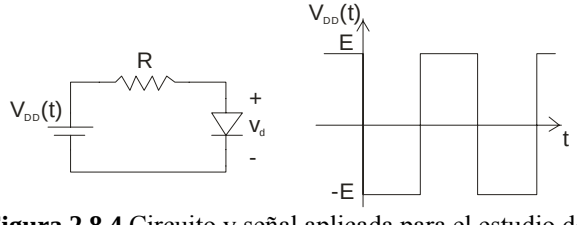


Figura 2.8.4 Circuito y señal aplicada para el estudio del transitorio de paso a inversa en una unión PN.

Antes de pasar de activa a inversa, se debe eliminar el exceso de huecos en la zona N para pasar a la situación correspondiente a la polarización en inversa. Mientras haya un exceso de huecos, la tensión v_D que cae en el diodo será positiva y la corriente vendrá dada por la

siguiente expresión (si la tensión de la fuente de alimentación es del orden de varios voltios):

$$I_D(t) = \frac{-E + v_d(t)}{R} \approx \frac{-E}{R}. \qquad (2.41)$$

Se denomina tiempo de almacenamiento (t_{sd}: *storage delay time*) al tiempo necesario para eliminar el exceso de carga y que v_d se anule. Una vez que eliminado el exceso de carga, se tiene que llegar a la situación correspondiente al estado en inversa. Para ello, se deben eliminar minoritarios y aumentar la tensión del diodo en inversa. La siguiente gráfica ilustra claramente los dos pasos del proceso:

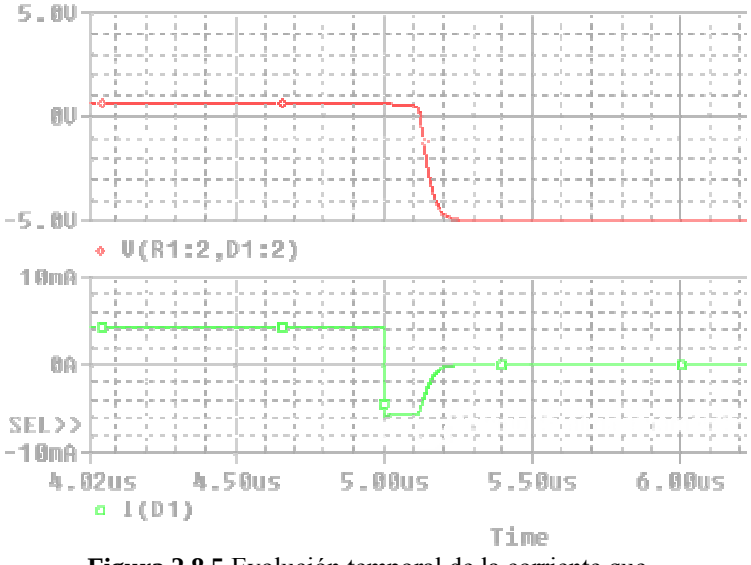


Figura 2.8.5 Evolución temporal de la corriente que atraviesa el diodo y de la tensión que cae en sus extremos durante un transitorio de paso a inversa (simulación realizada con Pspice).

Modelos de pequeña señal. Régimen AC

Introducción

Antes de explicar el modelo de pequeña señal de una unión PN, vamos a hacer un pequeño inciso para exponer qué es un modelo de pequeña señal y cuál es su utilidad

Comentario: ¿Qué es un modelo de pequeña señal?

Supongamos un circuito en el que se cumple una relación de proporcionalidad entre las señales de entrada y de salida (ya sean voltajes o corrientes):

$$y(x) = Ax, \tag{2.42}$$

donde A es una constante. Este circuito es lineal (su curva característica de transferencia está descrita por una recta) y se verifica el principio de superposición:

$$y(x_0 + \Delta x) = Ax_0 + A\Delta x = y(x_0) + y(\Delta x)\,. \tag{2.43}$$

En una defínción más general, se define un sistema lineal como aquel que verifica el principio de superposición, cualquiera que sea la relación entre la entrada y la salida (puede ser además de una relación de proporcionalidad, de tipo diferencial, integral o combinación lineal de las tres posibilidades).

El principio de superposición nos permite estudiar la respuesta de un circuito descomponiendo el problema en dos partes: primero se estudia la salida con ciertos valores fijos de las variables de entrada (polarización) y luego cómo cambia la respuesta del circuito al cambiar el valor de una de las variables de entrada.

En general, las curvas de transferencia de los dispositivos no son lineales (por ejemplo, la curva I-V de un diodo es de tipo exponencial) y no se puede aplicar el principio de superposición. Para evitar esto y conseguir trabajar con sistemas lineales, se linealiza la curva de transferencia en torno a un cierto punto de trabajo (o de polarización) desarrollando en serie de Taylor:

$$y(x_0 + \Delta x) = y(x_0) + \left(\frac{\partial y}{\partial x}\right)_{x=x_0} \Delta x\,. \tag{2.44}$$

Esta aproximación nos permite:

- *Poder seguir descomponiendo el problema en dos: respuesta en el punto de polarización ($y(x_0)$) y respuesta debido a variaciones en torno a este punto de polarización:*

$$\left(\left(\frac{\partial y}{\partial x}\right)_{x=x_0} \Delta x\right)$$

- *Obtener una relación de proporcionalidad entre las variaciones en la salida y las variaciones en la entrada:*

$$\Delta y = \left(\frac{\partial y}{\partial x}\right)_{x=x_0} \Delta x,\,. \tag{2.45}$$

donde: $\Delta y = y(x_0 + \Delta x_i) - y(x_0)$.

Se entiende por análisis de pequeña señal *el estudio de la respuesta del circuito ante las variaciones de tensión (o corriente) pequeñas (Δx) en torno a un punto de polarización DC (dado por (x_0,y_0)). Se impone que sean pequeñas para que se cumpla la relación de linealidad para que en el desarrollo en serie de Taylor* (2.44) *podamos quedarnos sólo con los dos primeros términos y, de este modo, poder aplicar el principio de superposición y obtener la relación lineal* (2.45) *entre las variaciones de la señal de salida y de salida.*

Por tanto, con un modelo de pequeña señal en primer lugar se resuelve el circuito en continua, para obtener el punto de polarización. Posteriormente, se anulan las fuentes que polarizan al circuito y se

estudia la respuesta de éste ante variaciones de los valores de estas fuentes. Posteriormente se particularizaremos este explicación para el caso de una unión PN y veremos un ejemplo de aplicación.

Cuando además el cambio en las señales de entrada en pequeña señal es de tipo sinusoidal, hablamos de régimen de alterna o AC.

Modelo de pequeña señal y baja frecuencia

En este apartado particularizaremos lo estudiado en el anterior para una unión PN. Supongamos la siguiente tensión aplicada al diodo[3]:

$$v_D(t) = V_D + v_d(t), \tag{2.46}$$

donde V_D es la tensión de polarización (DC) y $v_d(t)$ las variaciones de pequeña señal. En este apartado supondremos que son independientes del tiempo o lo suficientemente lentas como para que no influya el tiempo necesario para modificar la carga almacenada en las zonas neutras.

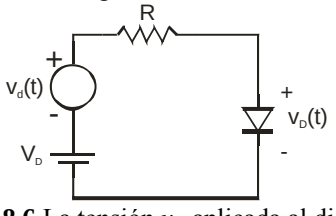


Figura 2.8.6 La tensión v_D aplicada al diodo se ha descompuesto en dos componentes: una componente continua que fija el punto de polarización del diodo (V_D) y una eventual variación en torno a dicho valor (v_d).

Veamos qué corriente, en función del tiempo, pasa por el diodo:

$$i_D(t) = I_S e^{v_D(t)/V_T} = I_S e^{(V_D + v_d(t))/V_T} = I_D e^{v_d(t)/V_T}, \tag{2.47}$$

donde

$$I_D = I_S e^{V_D/V_T} \tag{2.48}$$

Como ya sabíamos, no hay una relación de linealidad entre la corriente y la tensión en el diodo ni entre las variaciones de tensión y corriente en el diodo. Sin embargo, esto último puede solucionarse. Desarrollando en serie de Taylor podemos obtener una relación lineal entre la corriente y la tensión de pequeña señal:

[3]La notación aquí empleada es bastante habitual cuando se emplean modelos de pequeña señal. En minúscula y con subíndice en mayúscula se expresa la tensión (o corriente, en su caso) total. Cuando se escriben con mayúscula tanto la variable como el subíndice se indica un punto de polarización (señal no variable). Con minúscula y con subíndice en minúscula se representan las variaciones de pequeña señal.

$$i_D(t) = I_D + \left(\frac{\partial i_D}{\partial v_D}\right)_{v_D=V_D} v_d = I_D + i_d(t). \qquad (2.49)$$

Es decir:

$$i_d(t) = \left(\frac{\partial i_D}{\partial v_D}\right)_{v_D=V_D} v_d \equiv \frac{1}{r_d} v_d \qquad (2.50)$$

De este modo, se obtiene una relación lineal entre las pequeñas variaciones de corriente (i_d) ante pequeños cambios en la tensión aplicada al diodo.(v_d). El coeficiente de proporcionalidad se expresa mediante una resistencia (r_d), puesto que relaciona una tensión entre los terminales de un elemento (el diodo) con la corriente que circula por dicho elemento debido a esa tensión.

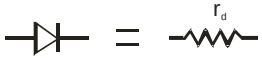


Figura 2.8.7 Modelo de pequeña señal y baja frecuencia de un diodo.

Observaciones:

- r_d depende del punto de polarización (DC)
- r_d^{-1} no coincide necesariamente con la pendiente de la recta del modelo de gran señal lineal a tramos, excepto si ésta se calcula precisamente a partir de la pendiente en el punto de polarización dado.
- En general, para resolver un ejercicio en el que intervienen variables de pequeña señal, se siguen los siguientes pasos:
  - Se resuelve el circuito en DC (se hacen las fuentes de pequeña señal igual a cero).
  - Se sustituye el elemento por su modelo de pequeña señal y se anulan las fuentes no variables que fijan el punto de polarización[4].
  - Se resuelve el circuito con las fuentes de pequeña señal para calcular las variaciones de pequeña señal en la salida.

Ejercicio:

Considere el circuito de la Figura 2.8.8. La fuente de alimentación V^+ tiene un valor en continua igual a 10V, sobre el que se superpone un rizado dado por una sinusoide de 60 Hz y 0.5 V de amplitud. Calcule el voltaje que cae en el diodo en continua y la amplitud de la señal sinusoidal que aparece en la salida superpuesta sobre el valor en continua.

[4]Anular una fuente de corriente equivale en un circuito a quitarla (dejando el circuito abierto) y una fuente de tensión a sustituirla por un cortocircuito.

Suponga que el diodo tiene una caída de 0.7 V con una corriente de 1 mA.

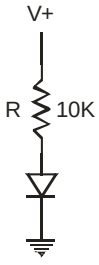


Figura 2.8.8 Circuito del ejercicio

Modelos de pequeña señal dependientes de la frecuencia

Introducción

Como se sabe, la capacidad está asociada a la presencia de cargas, que modifican la distribución de potencial acumulando energía. Cuando se altera el potencial, cambiará la carga almacenada. El tiempo necesario para ello dependerá de lo fácil que pueda circular o modificarse la carga (p.e., en un circuito RC, depende del valor de la resistencia. En una unión PN polarizada en directo, de lo rápido que se recombine el exceso de carga). Como es sabido, en los condensadores de láminas paralelas o, en general, en los sistemas en los que la carga depende linealmente del voltaje aplicado, la capacidad se define como:

$$C = \frac{Q}{V}. \tag{2.51}$$

Sin embargo, hay sistemas en los que la carga depende de forma más complicada del voltaje: $Q = Q(V)$. En este caso, se linealiza la función $Q(V)$ en torno a cierto punto de interés y se define la capacidad como:

$$C = \frac{dQ}{dV}, \tag{2.52}$$

de forma que el cambio en la carga almacenada se relaciona con el cambio en el voltaje según:

$$\Delta Q = C\Delta V. \tag{2.53}$$

En el diodo hay dos tipos de carga: fijas y móviles y darán lugar a dos capacidades en el modelo de pequeña señal.

Capacidad de unión o de transición

Se debe a la carga de los iones fijos de la zona de carga espacial. La carga Q almacenada en la z.c.e. de un diodo es:

$$|Q| = qAx_{p0}N_A = qAx_{n0}N_D = qA\frac{N_A N_D}{N_A + N_D}W \tag{2.54}$$

Sustituyendo el valor de W:

$$Q = A\left[2q\varepsilon(V_0 - v_D)\frac{N_A N_D}{N_A + N_D}\right]^{1/2} \tag{2.55}$$

Por tanto, la capacidad (C_j) asociada a esta carga es:

$$C_j = \frac{dQ}{dv_D} = \frac{A}{2}\left[\frac{2q\varepsilon}{(V_0 - v_D)}\frac{N_A N_D}{N_A + N_D}\right]^{1/2} \Rightarrow \boxed{C_j \equiv \frac{\varepsilon A}{W}} \tag{2.56}$$

Como puede verse, esta expresión es análoga a la de un condensador de láminas planoparalelas de área A y separadas una distancia W.

Observaciones:

- Si la unión no es abrupta, sino que el perfil de impurezas cambia de otra forma, la relación entre W y v_D es diferente. Para una unión en la que el perfil de impurezas cambia linealmente se cumple:

$$C_j = \frac{A}{2}\left[\frac{2q\varepsilon}{(V_0 - v_D)}\frac{N_A N_D}{N_A + N_D}\right]^{1/3}. \tag{2.57}$$

- Para una unión cualquiera, se obtendría:

$$C_j = \frac{A}{2}\left[\frac{2q\varepsilon}{(V_0 - v_D)}\frac{N_A N_D}{N_A + N_D}\right]^{m}, \tag{2.58}$$

  donde m es un exponente que depende del perfil de impurezas y que varía entre 1/2 (unión abrupta) y 1/3 (unión lineal).
- Normalmente se suele expresar la capacidad de unión C_j en función de su valor en ausencia de polarización:

$$C_j = \frac{A}{2}\left[\frac{2q\varepsilon}{V_0}\frac{N_A N_D}{N_A + N_D}\right]^{m}\frac{1}{(1 - \frac{v_D}{V_0})^m} \Rightarrow \boxed{C_j \equiv \frac{CJO}{(1 - \frac{v_D}{V_0})^m}} \tag{2.59}$$

- En inversa, la carga almacenada es mayor que en directa, pero la capacidad es menor. Además, cuanto mayor sea la tensión inversa aplicada, menor será la capacidad y mayor la carga fija en la z.c.e.
- Supongamos un diodo P $^+$ N. Entonces:

$$C_j = \frac{A}{2}\left[\frac{2q\varepsilon}{(V_0 - v_D)}N_D\right]^{m}. \tag{2.60}$$

  Por tanto, a partir de medidas experimentales de la capacidad C_j en función de la tensión podemos determinar la concentración de

impurezas donadoras (N_D) del diodo.

Capacidad de difusión o de almacenamiento

A diferencia de la anterior, se debe a la carga móvil almacenada en las zonas neutras de la unión. Supongamos de momento una unión P^+N. Ya sabemos que la carga debida al exceso de huecos en la zona P es:

$$Q_p = I_p \tau_p \approx I\tau_p = \tau_p I_S e^{v_D / V_T} \tag{2.61}$$

Por tanto, la capacidad asociada a pequeños cambios en Q_p es:

$$Cs = \frac{dQ_p}{dv_D} = \tau_p \frac{1}{V_T} I_S e^{v_D / V_T} \Rightarrow C_S = \frac{I\tau_p}{V_T} \tag{2.62}$$

Si consideramos una unión PN hay que tener en cuenta también la contribución de Q_n:

$$C_s = \frac{I_p \tau_p}{V_T} + \frac{I_n \tau_n}{V_T} \Rightarrow \boxed{C_s = \frac{I}{V_T}\tau_T} \tag{2.63}$$

Como vemos, en lugar de usar dos parámetros diferentes (τ_n y τ_p), se ha definido un único parámetro τ_T (cuyo valor está comprendido entre τ_n y τ_p) y que se denomina tiempo de tránsito.

Modelo de pequeña señal

Finalmente, el modelo de pequeña señal completo se obtiene al incluir tanto la influencia sobre la corriente de los incrementos de tensión de baja frecuencia (rigurosamente, de continua) como el efecto de ambas capacidades, que limita la velocidad a la que puede cambiar la tensión en los extremos del diodo.

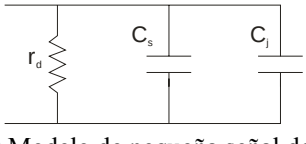


Figura 2.8.9 Modelo de pequeña señal de un diodo de unión.

2.9 Tipos de diodos y aplicaciones

En este apartado describiremos brevemente algunos tipos de diodos y sus principales aplicaciones.

Rectificadores

Se aprovecha la característica del diodo de dejar pasar la corriente sólo en un único sentido. Deben tener una característica lo más próxima a la de un diodo ideal:

- Muy baja corriente en inversa
- Tensión umbral próxima a 0 V
- $R_d = 0\ \Omega$
- Tensión de ruptura elevada

Ejemplos de diodos rectificadores comerciales: 1N4001, BYV95C.

Diodos de conmutación

Se usan en aplicaciones en las que se requiere que el tiempo de conmutación entre los estados de conducción y bloqueo de la corriente sea muy bajo.

Ejemplo: UF4001.

Diodos Zéner. Regulación de voltaje

Se usan para estabilizar (o limitar) la tensión en un determinado punto de un circuito. Ejemplo: eliminación del rizado de una fuente de alimentación.

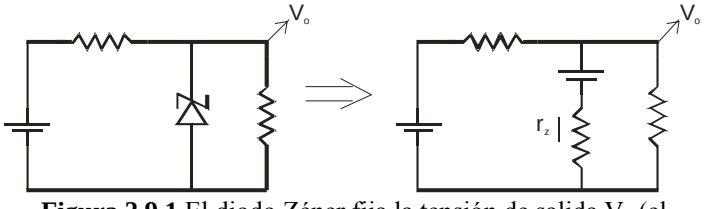


Figura 2.9.1 El diodo Zéner fija la tensión de salida V_0 (el valor de r_z es muy pequeño).

Varactores o diodos varicap

El símbolo de un diodo varicap se muestra en la siguiente figura:

Figura 2.9.2 Símbolo de un diodo varicap.

Estos diodos se usan para conseguir capacidades controladas por una tensión (por ejemplo, para su uso en filtros sintonizados) ya que la capacidad de unión depende del voltaje aplicado entre sus extremos (V_D). Efectivamente, recordemos que:

$$C_j \equiv \frac{CJO}{(1-\frac{V_D}{V_0})^m} \propto \left|V_D^{-m}\right| \text{ (en inversa)}. \qquad (2.64)$$

Diodos emisores de luz (LED)

El símbolo de estos diodos se muestra en la Figura 2.9.3.

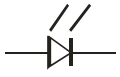


Figura 2.9.3 Símbolo de un diodo emisor de luz

En el arseniuro de galio y otros semiconductores compuestos, la recombinación de los minoritarios da lugar a la emisión de fotones. Variando la composición de algunos semiconductores (como $GaAs_{1-x}\ P_x$), podemos modificar la anchura de la banda prohibida y, por tanto, la longitud de onda de los fotones emitidos. La intensidad de la luz es proporcional a la intensidad de corriente que circula por el diodo. Por tanto, podemos establecer la intensidad luminosa polarizando adecuadamente el diodo.

Fotodiodos o diodos fotodetectores

Su símbolo es:

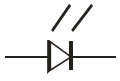


Figura 2.9.4 Símbolo de un diodo fotodetector

Al contrario que los LED, se basan en el mecanismo de generación. La luz incidente rompe un enlace covalente y genera un par electrón-hueco. Los fotodiodos se polarizan en inversa. En ausencia de luz, la corriente es despreciable ($-I_S$). Sin embargo, al iluminar el diodo, se genera una corriente que, al igual que la corriente inversa de saturación, es independiente de la tensión inversa pero que, en este caso, depende de la intensidad de la iluminación.

2.10 Unión metal-semiconductor

En este apartado veremos las uniones entre un metal y un semiconductor y cómo se pueden obtener características similares a las de un diodo de unión PN. Más importante aún, veremos que además pueden lograrse uniones entre metales y semiconductores no rectificadoras, es decir, que permiten el paso de la corriente en ambos sentidos. Esto es importante porque en la fabricación de dispositivos semiconductores siempre es necesario establecer contactos metálicos hacia las conexiones externas y, en principio, no es admisible la limitación de que conduzcan sólo en un sentido.

En un diodo de unión PN hemos visto que la diferencia en la posición del nivel de Fermi en los dos semiconductores provoca que al unirlos se desplacen los portadores para igualar el nivel de Fermi en toda la estructura, apareciendo una zona de carga espacial, caracterizada por la curvatura de las bandas (y el consiguiente campo eléctrico) y la presencia de cargas fijas debidas a las impurezas ionizadas.

En una unión metal-semiconductor sucede algo similar. La diferencia en la posición del nivel de Fermi provoca el movimiento de los electrones al unir ambos metales hasta que el nivel de Fermi se hace uniforme en toda la estructura.

En el caso de la unión PN, ambos semiconductores poseen la misma estructura de bandas y el nivel de Fermi en cada semiconductor puede fijarse, por ejemplo, respecto a la banda de conducción del mismo. Sin embargo, el metal y el semiconductor tienen diferente estructura de bandas y, por tanto, la banda de conducción no es equivalente en ambos materiales. Sin embargo, necesitamos comparar la posición del nivel de Fermi en ambos materiales. Para tener un punto de referencia común, elegimos la energía del vacío (E_0), definida como la que tendría un electrón sin energía cinética no ligado a ninguno de estos materiales.

Se definen la función trabajo como la energía que habría que aportar a un electrón para que pase de tener la energía correspondiente al nivel de Fermi a tener la del nivel del vacío. Para el metal y el semiconductor, respectivamente:

$$q\phi_m = E_0 - E_{Fm} \tag{2.65}$$

$$q\phi_s = E_0 - E_{Fs} \tag{2.66}$$

Además, en el caso del semiconductor, se define la afinidad electrónica como la energía que habría que aportar a un electrón situado en el fondo de la banda de conducción para que escapase del material y pasase al nivel del vacío E_0. Esto es, la afinidad electrónica del semiconductor viene dada por:

$$q\chi = E_0 - E_c \tag{2.67}$$

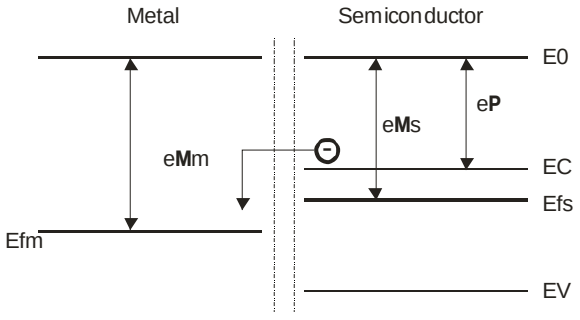


Figura 2.10.1 Funciones trabajo del metal y del semiconductor y afinidad electrónica. Los diagramas de bandas se corresponden con la situación en la que los dos materiales no están en contacto.

En la Figura 2.10.1 se muestran los diagramas de bandas del metal y de un semiconductor (tipo N) aislados y el significado gráfico de las magnitudes anteriormente definidas. En el caso representado, la función trabajo del metal es mayor que la del semiconductor ($\boldsymbol{M}_m > \boldsymbol{M}_s$). Esto implica que cuando los materiales se pongan en contacto, habrá un flujo de electrones del semiconductor al metal, hasta que se igual el nivel de Fermi.

Al igual que sucedía en la unión PN, este traspaso de electrones del semiconductor al metal no puede continuar indefinidamente, sino que se corta debido al campo eléctrico opuesto a este movimiento que se crea por las impurezas positivas descompensadas al irse los electrones. La misma carga que aparece en el semiconductor, se produce también en el metal, pero de signo contrario. No obstante, al ser la conductividad del metal muy grande, el campo eléctrico en su interior debe ser nulo y toda la carga negativa se acumula en la interfaz con el semiconductor, no existiendo por tanto una zona de carga espacial en la región correspondiente al metal.

La Figura 2.10.2 muestra el diagrama de bandas de la estructura metal-semiconductor en equilibrio térmico. En la banda de conducción del semiconductor aparece una barrera de potencial que dificulta el tránsito de electrones del semiconductor hacia el metal. La altura de esta barrera de potencial viene dada por la diferencia de los potenciales de las funciones trabajo de ambos materiales:

$$V_0 = \phi_m - \phi_s \qquad (2.68)$$

Obsérvese que debido a que para las energías en torno al nivel de Fermi no hay estados energéticos que puedan ser ocupados en el semiconductor, aparece también una barrera de potencial del metal al semiconductor. Esto es consecuencia directa de la discontinuidad en las bandas de conducción de ambos materiales y por eso no sucede en la unión PN, en la que ambos materiales tienen la misma banda de conducción. La altura de esta barrera viene dada por:

$$\phi_B = \phi_m - \chi \tag{2.69}$$

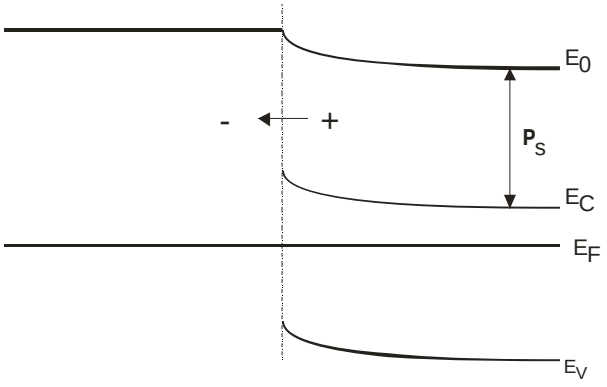


Figura 2.10.2 Unión metal-semiconductor. La gráfica muestra la curvatura de la banda de conducción del semiconductor y la barrera de potencial que dificulta el tránsito de los electrones del semiconductor hacia el metal.

Si ahora aplicamos una tensión positiva en el metal respecto del semiconductor, atraeremos electrones del semiconductor hacia el metal. O lo que es lo mismo, disminuye la barrera de potencial V_0, rompiéndose la situación de equilibrio y produciéndose corriente neta. La corriente será tanto mayor cuanto mayor sea el voltaje aplicado y la consiguiente disminución de la barrera. Si, por el contrario, se aplica una tensión negativa, se incrementa esta barrera de potencial, impidiendo el flujo de electrones del semiconductor hacia el metal. Además, la barrera $\boldsymbol{N}_B$ entre el metal y el semiconductor impide el tránsito de electrones del metal al semiconductor. Por tanto, cuando la tensión aplicada es negativa, no circula corriente por la unión.

Este comportamiento rectificador es completamente análogo al de la unión PN y la curvas características I-V son similares en ambos casos.

En el caso de una unión entre un metal y un semiconductor tipo P se obtiene el mismo comportamiento rectificador si $\boldsymbol{N}_m < \boldsymbol{N}_s$. En este caso se forma una barrera en la banda de valencia que dificulta el paso de huecos del semiconductor al metal.

Pueden lograrse uniones no rectificadoras (esto es, que permitan el paso de la corriente en ambos sentidos) si $\boldsymbol{N}_m < \boldsymbol{N}_s$ y el semiconductor es tipo N o si $\boldsymbol{N}_m > \boldsymbol{N}_s$ y el semiconductor es tipo P. En estos casos, la carga que aparece en el semiconductor no la proporciona la deplexión del mismo, sino la acumulación de mayoritarios provenientes del metal. Por ejemplo, en el primer caso, el nivel de Fermi en el metal es mayor que en el semiconductor y por tanto, para igualar el nivel de Fermi, hay un transvase de electrones del metal al semiconductor y la barrera que se

forma cuando se alcanza el equilibrio térmico es precisamente en esta dirección (ver Figura 2.10.3).

Si ahora aplicamos una tensión positiva en el metal respecto del semiconductor, los electrones no tienen ningún problema para ir hacia el metal, produciéndose corriente. Si la tensión es en sentido contrario, los electrones que van a pasar del metal al semiconductor ven una barrera de potencial, pero esta es mucho menor que en los contactos rectificadores (E_c - E_F), por lo que una pequeña disminución de la barrera causada por la tensión negativa aplicada es suficiente también para permitir el paso de corriente. Por tanto, en este tipo de uniones metal-semiconductor se permite el paso de la corriente en ambos sentidos, por lo que se denominan no rectificadoras o contactos óhmicos.

También pueden lograrse contactos óhmicos dopando mucho el semiconductor, para conseguir una zona de deplexión delgada. De este modo, en el caso en el que exista una barrera de potencial, ésta será muy estrecha y podrá ser atravesada mediante efecto túnel.

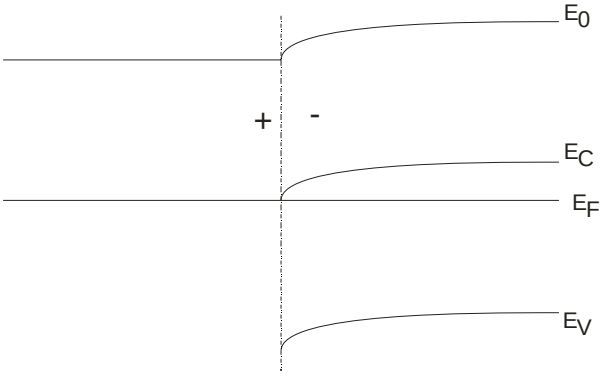


Figura 2.10.3 Diagrama de bandas de una unión metal-semiconductor no rectificadora en equilibrio térmico.

2.11 Heterouniones

Ya hemos estudiado las uniones de dos semiconductores (P y N) del mismo material y las uniones entre un metal y un semiconductor. En este apartado estudiaremos las uniones entre dos materiales semiconductores diferentes.

Por lo general, estos dos semiconductores tendrán diferentes anchuras de la banda prohibida, afinidades electrónicas y funciones trabajo. Esto provoca que estas uniones, además de la curvatura de las bandas (barrera de potencial) debida a la redistribución de cargas para igualar los niveles de Fermi, contengan discontinuidades en las bandas de conducción y de valencia y que las barreras para los electrones y para los huecos no sean las mismas. En principio cabe esperar que la

discontinuidad en la banda de conducción dé cuenta de la diferencia de afinidad electrónica entre los dos materiales (ΔEc = q(χ_{s1} - χ_{s2})). La discontinuidad en la banda de valencia sería entonces la resta de ΔEc con la diferencia de las anchuras de la banda prohibida (ΔE_g) en ambos materiales (ver Figura 2.11.1).

Como puede observarse, en una heterounión las barreras de potencial que tienen que superar los electrones y huecos para pasar de un lado de la unión al otro son diferentes. Esto permite que con las heterouniones se pueda controlar la proporción de electrones y huecos inyectados en la unión. Esta propiedad se emplea, por ejemplo, en los transistores bipolares de heterounión (ó HBT, según las siglas inglesas) en los cuales se aumenta la proporción de electrones inyectados a través de una unión respecto de la de huecos sin tener que ajustar los dopados de los semiconductores (que se fijan entonces de acuerdo con otros requerimientos). Esto permite mejorar la eficiencia de estos transistores, como se comprobará en el capítulo 3.

Por otro lado, las heterouniones también se emplean en transistores de efecto campo. Como puede verse en la heterounión mostrada en la Figura 2.11.1 se ha formado un pozo de potencial en donde los huecos pueden quedar confinados. Análogamente, en las heterouniones entre AlGaAs y GaAs se forma un pozo de potencial en la banda de conducción que puede confinar a los electrones, formando un gas bidimensional (los electrones sólo pueden moverse en las dimensiones del plano formado por la interfaz) que tiene buenas propiedades para la conducción.

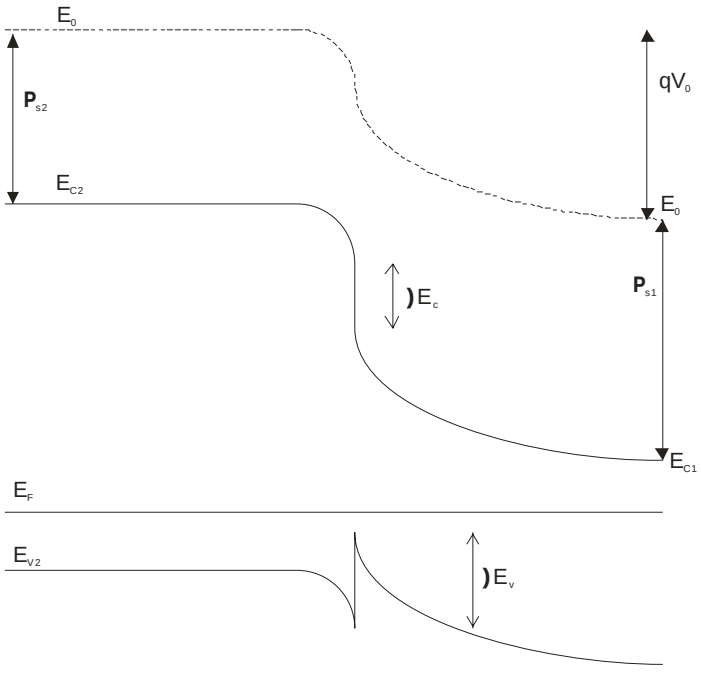


Figura 2.11.1 Diagrama de bandas de una heterounión, en equilibrio térmico, formada por un semiconductor tipo N y un semiconductor tipo P con diferente ancho de banda prohibida.

REFERENCIAS

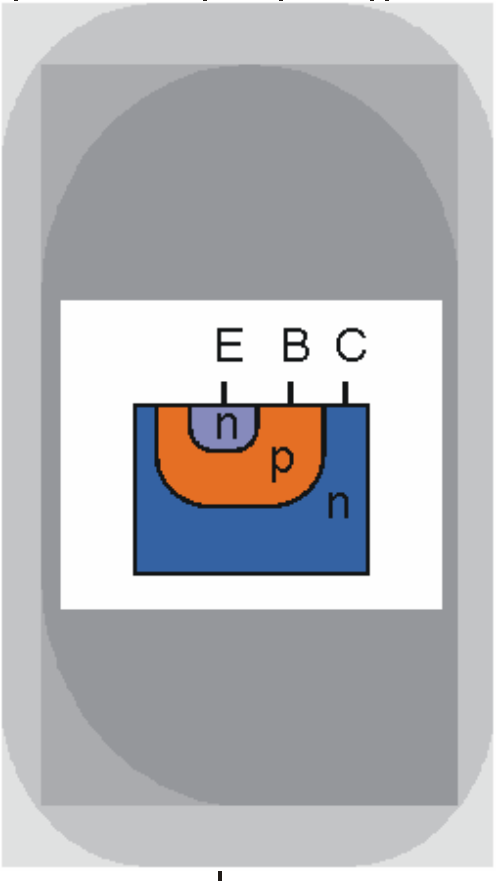
[1] B.G. Streetman, S. Banerjee, *Solid State Electronic Devices*, Prentice Hall, 2000.

[2] S. Dimitrijev, *Understanding Semiconductor Devices*, Oxford University Press, 2000.

Capítulo 3

El Transistor Bipolar De Unión (BJT)

Transistor Bipolar



Índice

Objetivos

- Describir el transistor bipolar.
- Entender el funcionamiento básico del transistor bipolar.
- Describir los distintos modos de operación de este dispositivo.
- Evaluar de forma cuantitativa la corriente que circula por los diferentes terminales del transistor.
- Proporcionar modelos eléctricos equivalentes en gran señal y pequeña señal. Los modelos deberán ser lineales y permitirán analizar o diseñar circuitos electrónicos que incluyan a este dispositivo.
- Simplificar los modelos equivalentes encontrados para facilitar el análisis de estos circuitos.
- Estudiar la conmutación de este dispositivo entre sus estados.
- Analizar la influencia de efectos de segundo orden no considerados en un primer análisis simplificado del dispositivo.
-

Palabras Clave

Transistor bipolar de unión.
Transistor npn y pnp.
Emisor.
Base.
Colector.
Ganancia de corriente en base común y en emisor común.
Activa.
Saturación.
Activa inversa.
Corte.
Modulación de la anchura de la base.
Tensión Early.
Ruptura.
Deriva en la base.
Ecuaciones de Ebers-Moll.
Efecto Kirk.
Modelo de gran señal.
Modelo de pequeña señal.
Transconductancia.
Resistencia de salida.
Capacidades de las uniones.
Resistencias parásitas.
Respuesta en frecuencia.

3.1 Introducción

El transistor bipolar fue el primer dispositivo activo de estado sólido. Fue inventado en 1949 en los Laboratorios Bell por W. Schockley, J. Bardeen y W. Brattain (que recibieron el premio Nobel en 1956). También se suele denominar por sus siglas inglesas BJT (*bipolar junction transistor*).

Se trata de un dispositivo formado por dos uniones y que tiene tres terminales (llamados emisor, base y colector). Hay dos tipos, npn y pnp:

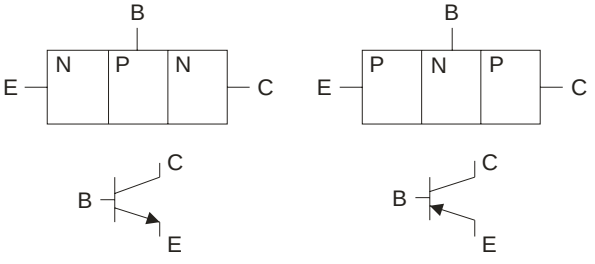


Figura 3.1.1 Estructura y símbolo de un transistor bipolar npn (izquierda) y pnp (derecha)

Entre sus principales aplicaciones podemos distinguir:

- Analógicas: amplificadores, seguidores de tensión, ...
- Digitales: conmutadores

3.2 Fundamentos básicos. Descripción cualitativa.

El BJT como fuente de corriente controlada por voltaje

En este apartado vamos a introducir el BJT como una fuente de corriente controlada por tensión. Supongamos una unión PN polarizada en inverso. Se puede considerar que es una fuente de corriente casi ideal porque la corriente que la atraviesa es independiente de la tensión entre sus extremos, como se ilustra en la Figura 3.2.1.

Sin embargo, presenta un inconveniente: la corriente es muy pequeña
(I_S) y está limitada por la generación térmica de minoritarios en las cercanías de la unión. Esta corriente podría, no obstante, incrementarse generando minoritarios, por ejemplo, mediante luz (ver Figura 3.2.2). Además, con la intensidad de la luz podemos controlar la intensidad de la fuente de corriente. Sería bueno poder hacer esto eléctricamente. Para

ello, podríamos añadir una unión más al sistema, puesto que en una unión P^+N se inyectan huecos desde la zona P^+ en la zona N y el número de huecos inyectados depende de la tensión aplicada en esta unión. Por tanto, se tiene entonces una fuente de corriente controlada por tensión (que determina el número de huecos inyectados en el semiconductor N), como se observa en la Figura 3.2.3.

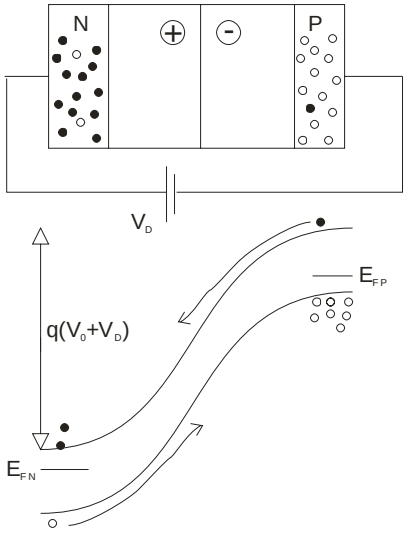


Figura 3.2.1 Una unión PN en inverso se comporta como una fuente de corriente: la corriente no depende de la tensión en inverso aplicada, sino de lo rápido que se reponen los minoritarios que caen por la barrera de potencial.

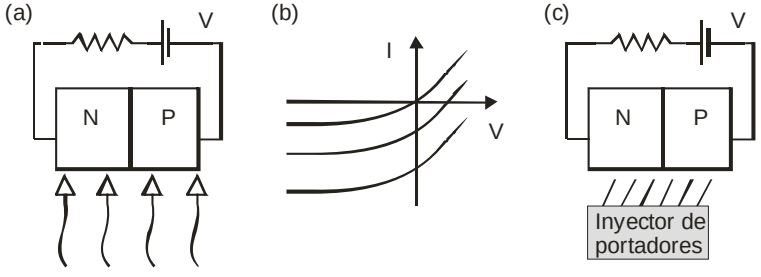


Figura 3.2.2 El incremento de la concentración de minoritarios puede realizarse con luz (a) y provoca el aumento de la corriente inversa (b). Sería deseable contar con otro procedimiento de inyección de minoritarios (c) que pudiésemos controlar eléctricamente.

Las corrientes en el BJT

En un transistor bipolar pnp polarizado como en la Figura 3.2.3 se producen los siguientes flujos de portadores:

- 1: huecos inyectados que se recombinan sin llegar al colector
- 2: huecos inyectados en el colector
- 3: corriente inversa de saturación en la unión PN de la base y el colector
- 4: electrones suministrados por el contacto de base para recombinarse con los huecos en la base
- 5: electrones inyectados en el emisor debido a que el emisor y la base forman una unión PN polarizada en directo

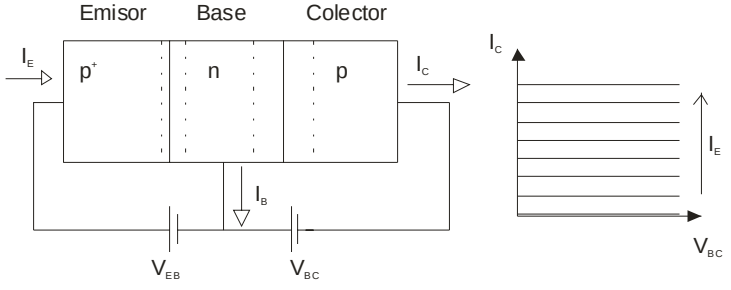


Figura 3.2.3 Polarización en activa de un BJT pnp

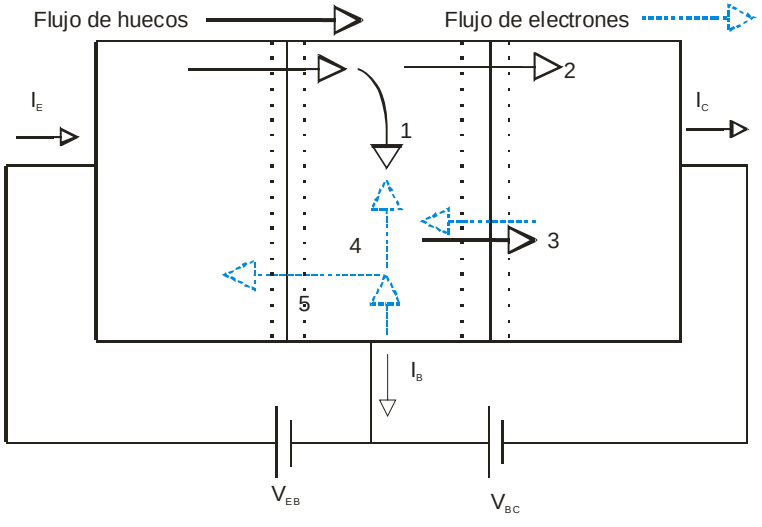


Figura 3.2.4 Flujo de portadores en un transistor pnp polarizado en activa

En un buen transistor es deseable que se verifiquen las siguientes condiciones:

- Casi todos los huecos aportados por el emisor llegan al colector.

Por tanto, se requiere que exista poca recombinación en la base, lo cual se consigue haciendo que la base sea estrecha ($W_b << L_p$). Se define el *factor de transporte*:

$$B \equiv \alpha_T \equiv \frac{I_{Cp}}{I_{Ep}} \approx \frac{I_C}{I_{Ep}}. \qquad (3.1)$$

(Por tanto, se busca que B sea aproximadamente 1).

- La corriente de emisor se debe casi exclusivamente a los huecos inyectados, no a los electrones procedentes de la base (para ello, hay que dopar el emisor mucho más que la base). Se define la *eficiencia de emisor* como:

$$\gamma \equiv \frac{I_{Ep}}{I_E} = \frac{I_{Ep}}{I_{Ep} + I_{En}}. \qquad (3.2)$$

(Por lo que en un buen transistor, se debe cumplir que γ sea aproximadamente igual a 1).

Calculemos ahora cuál es la relación existente entre las corrientes de colector y emisor (ganancia en base común):

$$\alpha_F \equiv \frac{I_C}{I_E} \approx \frac{I_{Cp}}{I_{Ep} + I_{En}} = \frac{I_{Cp}}{I_{Ep}} \frac{I_{Ep}}{I_{Ep} + I_{En}} = B \cdot \gamma. \qquad (3.3)$$

En un buen transistor se cumple que α_F es casi 1, puesto que el factor de transporte y la eficiencia de emisor son próximos a 1.

Por otra parte, la relación entre las corrientes de colector y base (ganancia en emisor común):

$$\beta_F \equiv \frac{I_C}{I_B} = \frac{I_C}{I_E - I_C} = \frac{1}{\frac{I_E}{I_C} - 1} = \frac{1}{\alpha_F^{-1} - 1} = \frac{\alpha_F}{1 - \alpha_F} \qquad (3.4)$$

De donde también se deduce que:

$$a_F = \frac{\beta_F}{\beta_F + 1}. \qquad (3.5)$$

Cuanto más próximo a 1 sea el valor de α_F mayor será la ganancia en corriente β_F. Valores típicos de β_F son 200 para transistores de baja potencia y 50 para transistores de potencia.

Modos de operación del BJT

Puesto que el BJT tiene cuatro uniones, puede operar según 4 modos diferentes, según estas uniones estén en directa o en inversa. Estas regiones de operación son:

Activa

Es el modo anteriormente analizado:

- unión BE en directo: $|V_{BE}| \cong 0.7$ V
- unión BC en inverso

En esta región de operación se cumple que:

$$\begin{aligned} I_C &= \alpha_F I_E, \\ I_C &= \beta_F I_B. \end{aligned} \tag{3.6}$$

Se usa sobre todo en aplicaciones analógicas (por ejemplo, como amplificador).

Corte

Las dos uniones están en inverso, por lo que no hay inyección de portadores del emisor al colector y todas las corrientes son prácticamente nulas.

Saturación

En esta región de operación:

- Unión BE en directo: $|V_{BE}| \cong 0.7$ V
- Unión BC en directo: $|V_{BE}| \cong 0.8$ V

En estas condiciones existe inyección de portadores desde el emisor, pero también desde el colector y, por tanto, se verifica (como luego comprobaremos con más detalle) que:

$$\begin{aligned} I_C &< \alpha_F I_E, \\ I_C &< \beta_F I_B. \end{aligned} \tag{3.7}$$

Además, se cumple que la diferencia de tensión entre el emisor y el colector es baja (del orden de 0.1-0.2 V). Por ejemplo, para un pnp:

$$V_{EC} = V_{EB} + V_{BC} \approx (0.7 - 0.6)\ \text{V} = 0.1\ \text{V} \tag{3.8}$$

Los modos de corte y saturación se emplean sobre todo en aplicaciones digitales.

Activa inversa

Este modo de operación es parecido al de la región activa, pero intercambiando los papeles de colector y del emisor. Es decir:

- Unión BE en inverso
- Unión BC en directo: $|V_{BC}| \cong 0.6$ V

Normalmente el transistor no opera en esta región. Recuérdese que el BJT no es completamente simétrico y está optimizado para trabajar en modo activo normal (por ejemplo, el emisor se dopa más que el colector).

3.3 Cálculo de las corrientes. Ecuaciones de Ebers-Moll

Introducción

Hasta ahora, hemos visto cualitativamente el funcionamiento del BJT, calculando unas corrientes a partir de otras a través de los parámetros α_F y β_F. Pero no sabemos el valor de estos parámetros ni cómo relacionar las corrientes con las tensiones aplicadas. Esto es lo que se pretende hacer en este apartado.

De forma análoga a como se hizo en el caso del diodo, se puede calcular la corriente de emisor (debida a los huecos) a partir de la corriente de difusión en la base justo en el límite de la z.c.e. entre el emisor y la base (se desprecia la recombinación en la z.c.e). Del mismo modo, la corriente de colector se calcula a partir de la corriente de difusión en el límite de la base con la z.c.e. espacial que comparte con el colector:

$$\begin{aligned} I_{Ep} &= -qAD_p \frac{d\delta p_n(x_n = 0)}{dx_n} \\ I_{Cp} &= -qAD_p \frac{d\delta p_n(x_n = W_b)}{dx_n} \end{aligned} \qquad (3.9)$$

Por tanto, lo primero que hay que hacer es resolver la ecuación de difusión para calcular el perfil de minoritarios inyectados en la base (δp_n):

$$\frac{d^2\delta p(x_n)}{dx_n^2} = \frac{\delta p_n(x_n)}{L_p^2} \qquad (3.10)$$

Solución de la ecuación de difusión en la base

En activa, el emisor inyecta huecos de forma que se genera un exceso en la base, justo en el borde de la z.c.e., igual a:

$$\Delta p_E = \delta p_n(0) = p_{n0}(e^{V_{EB}/V_T} - 1) \approx p_n e^{V_{EB}/V_T}. \qquad (3.11)$$

En el otro borde de la base, los huecos son barridos por el campo eléctrico hacia el colector, de modo que:

$$\Delta p_C = \delta p_n(W_b) = -p_{n0}. \qquad (3.12)$$

Si no hubiese corriente de base, el gradiente $\frac{d\delta p_n}{dx_n}$ sería el mismo en el colector y el emisor (para tener la misma corriente, como indican las ecuaciones (3.9). Por tanto, la solución de la ecuación de difusión sería una línea recta, como se indica en la figura inferior.

Sin embargo, la corriente de base no es nula, aunque sí pequeña, por lo que la solución real no es exactamente una línea recta, aunque se aproxima bastante (por debajo, para obtener menos gradiente en el lado del colector):

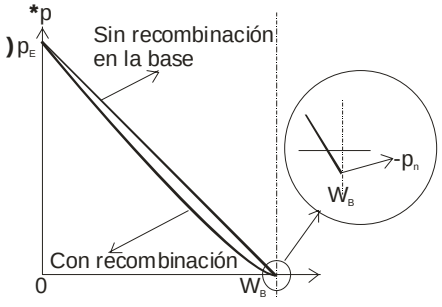


Figura 3.3.1 Exceso de minoritarios en la base en función de la posición (activa)

En activa inversa, se obtiene una solución completamente análoga, pero intercambiando los papeles del emisor y del colector.

Como la ecuación de difusión es lineal, la solución general (con inyección de huecos por parte del emisor y del colector) puede expresarse como la superposición de las soluciones obtenidas en activa (inyección sólo por parte del emisor) y activa inversa (inyección sólo por parte del colector). Esto se ilustra en la siguiente figura:

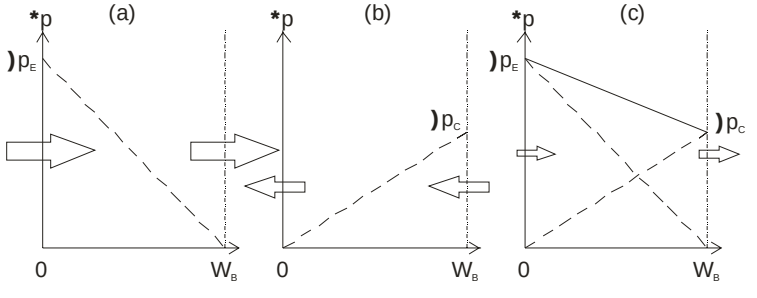


Figura 3.3.2 Concentración y flujo de minoritarios en la base (c) como superposición de la solución de la ecuación de difusión en activa (a) y en activa inversa (b)

Puede comprobarse fácilmente que la solución de la ecuación de difusión (3.10) es:

$$\delta p(x_n) = \Delta p_E \frac{e^{(W_b - x_n)/L_p} - e^{-(W_b - x_n)/L_p}}{e^{W_b/L_p} - e^{-W_b/L_p}} + \\ + \Delta p_C \frac{e^{x_n/L_p} - e^{-x_n/L_p}}{e^{W_b/L_p} - e^{-W_b/L_p}}. \tag{3.13}$$

Efectivamente se observa que la solución es la suma de dos términos con la misma forma, cada uno debido a la contribución de una de las uniones.

Evaluación de las corrientes

Conocido $\delta p(x_n)$, podemos calcular ya las corrientes (). Al realizar las derivadas, se obtendrá que, como sucedía con la concentración de minoritarios en la base, las corrientes también son superposición de la contribución de ambas uniones.

Suponiendo que $I_E = I_C$ (esto es, que la eficiencia de emisor γ es igual a 1) se obtiene:

$$\begin{aligned} I_E &= A\left(e^{V_{EB}/V_T} - 1\right) - B\left(e^{V_{CB}/V_T} - 1\right), \\ I_C &= B\left(e^{V_{EB}/V_T} - 1\right) - A\left(e^{V_{CB}/V_T} - 1\right), \\ I_B &= I_E - I_C, \end{aligned} \tag{3.14}$$

con:

$$\begin{aligned} A &= \frac{qAD_p}{L_p} \coth\left(\frac{W_b}{L_p}\right) p_{n0} \\ B &= \frac{qAD_p}{L_p} \cosh\left(\frac{W_b}{L_p}\right) p_{n0} \end{aligned} \tag{3.15}$$

El modelo de diodos acoplados. Ecuaciones de Ebers-Moll

Introducción

Las ecuaciones vistas en el apartado anterior (dependientes de la geometría) sólo son válidas para ese caso concreto de geometría uniforme y simple y suponiendo, además, $\gamma = 1$. En este apartado pretendemos generalizarlas, buscando expresiones adecuadas para cualquier geometría y dependientes de parámetros fácilmente medibles.

Para ello, nos basaremos en las ecuaciones y conclusiones del apartado anterior: la inyección de portadores en la base se puede descomponer en dos contribuciones debidas a dos diodos independientes.

Pero además de la corriente de emisor o colector debida a estos diodos, debemos superponer la contribución de los inyectados por el otro electrodo (colector o emisor) en la base y que llegan, respectivamente, al emisor o al colector.

Esta descripción se corresponde con el siguiente circuito equivalente:

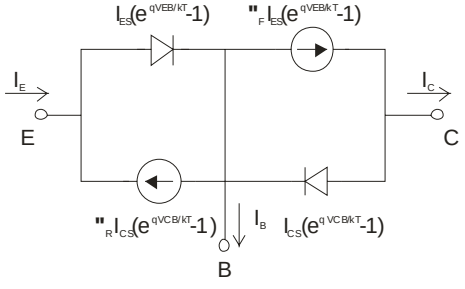


Figura 3.3.3 Circuito equivalente de un BJT correspondiente a las ecuaciones de Ebers-Moll (modelo de diodos acoplados)

Vemos pues que este modelo está descrito por la corriente de dos diodos independientes y de dos fuentes de corriente que dan cuenta del acoplamiento entre las dos uniones de la estructura.

Las ecuaciones de Ebers-Moll son las expresiones que relacionan las corrientes y tensiones en el BJT de acuerdo con este modelo de diodos acoplados.

Ecuaciones de Ebers-Moll

Según el modelo de diodos acoplados presentado anteriormente, las corrientes en los tres terminales de un BJT vienen dadas por las siguientes expresiones (conocidas como *ecuaciones de Ebers-Moll*):

$$\begin{aligned} I_E &= I_{ES}\left(e^{V_{EB}/V_T}-1\right)-\alpha_R I_{CS}\left(e^{V_{CB}/V_T}-1\right) \\ I_C &= \alpha_F I_{ES}\left(e^{V_{EB}/V_T}-1\right)-I_{CS}\left(e^{V_{CB}/V_T}-1\right) \\ I_B &= I_E - I_C \end{aligned} \qquad (3.16)$$

Estas ecuaciones relacionan las corrientes en el BJT con las tensiones aplicadas en sus uniones y son dependientes de cuatro parámetros (las corrientes inversas de saturación, I_{ES} e I_{CS} y los coeficientes de acoplamiento, α_F y α_R.

Observaciones:

- Con el convenio de signos seguido en este trabajo, en el caso de un transistor npn las ecuaciones son exactamente iguales que las de un pnp. Sólo hay que cambiar las tensiones de signo o, lo que es lo mismo, sustituir V_{BB} por V_{BE} y V_{CB} por V_{BC}.
- En muchos textos se definen como positivas las corrientes entrantes en el dispositivo. Con ese convenio de signos, habría que cambiar el signo en los sumandos del miembro de la derecha de la ecuación correspondiente a I_C y $I_B = -I_E - I_C$.
- Con $V_{CB} = 0$ V (y, en general, en la región activa) se cumple:

$$\begin{aligned} I_E &= I_{ES}\left(e^{V_{EB}/V_T} - 1\right) \\ I_C &= \alpha_F I_{ES}\left(e^{V_{EB}/V_T} - 1\right) \\ I_B &= (1-\alpha_F) I_{ES}\left(e^{V_{EB}/V_T} - 1\right) \end{aligned} \qquad (3.17)$$

Por tanto, tal y como habíamos definido en la introducción de este tema, se cumple que (en *activa*):

$$\begin{aligned} \frac{I_C}{I_E} &= \alpha_F \\ \frac{I_C}{I_B} &= \frac{\alpha_F}{1-\alpha_F} \equiv \beta_F \end{aligned} \qquad (3.18)$$

- Como puede verse, si cortocircuitamos la base y el colector, el BJT se comporta como un diodo.
- Análogamente, en la región activa inversa se cumple que:

$$\left.\frac{I_E}{I_C}\right)_{\text{activa inversa}} = \alpha_R \qquad (3.19)$$

- Como se ha comentado anteriormente, las ecuaciones de Ebers-Moll son dependientes de cuatro parámetros. Sin embargo, puede demostrarse que se cumple la siguiente relación entre ellos:

$$\alpha_F I_{ES} = \alpha_R I_{CS}. \qquad (3.20)$$

- Por tanto, sólo quedan tres parámetros independientes.

Otros modelos DC

Versión de transporte de las ecuaciones de Ebers-Moll

Basándonos en la relación (3.20) podemos expresar las ecuaciones de Ebers-Moll como:

$$
\begin{aligned}
I_E &= \frac{I_S}{\alpha_F}\left(e^{V_{EB}/V_T}-1\right) - I_S\left(e^{V_{CB}/V_T}-1\right) \\
I_C &= I_S\left(e^{V_{EB}/V_T}-1\right) - \frac{I_{CS}}{\alpha_R}\left(e^{V_{CB}/V_T}-1\right) \\
I_B &= I_E - I_C
\end{aligned}
\tag{3.21}
$$

Estas expresiones se conocen como la versión de transporte de las ecuaciones de Ebers-Moll.

En estas ecuaciones cambia el punto de vista respecto de las ecuaciones de Ebers-Moll. En lugar de expresar la corriente de las fuentes de corriente en función de la corriente que pasa por los diodos (mediante los coeficientes α_F y α_R), expresamos la corriente que pasa por los diodos en función de la de las fuentes de corriente.

De acuerdo con esto, se definen las corrientes de las fuentes de corriente como:

$$
\begin{aligned}
I_{EC} &= I_S\left(e^{V_{CB}/V_T}-1\right) \\
I_{CC} &= I_S\left(e^{V_{EB}/V_T}-1\right)
\end{aligned}
\tag{3.22}
$$

Y el circuito equivalente queda del siguiente modo:

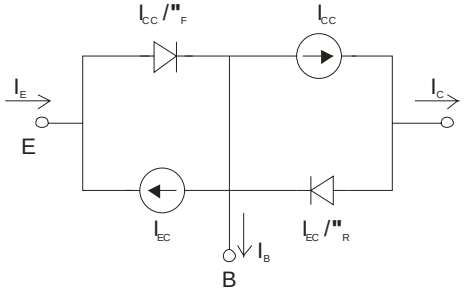


Figura 3.3.4 Versión de transporte del modelo de diodos acoplados

Versión de Spice

Puesto que las ecuaciones de Ebers-Moll tienen sólo tres parámetros independientes, se puede buscar un modelo en el que sólo aparezcan tres elementos en lugar de cuatro.

El modelo de Spice conecta directamente el emisor con el colector. Para obtenerlo, primero se calcula la corriente de base a partir de la figura anterior:

$$I_B = \frac{I_{CC}}{\alpha_F} + \frac{I_{EC}}{\alpha_R} - I_{EC} - I_{CC} = \\ = I_{EC}\left(\frac{1}{\alpha_R} - 1\right) + I_{CC}\left(\frac{1}{\alpha_F} - 1\right) = \frac{I_{EC}}{\beta_R} + \frac{I_{CC}}{\beta_F} \tag{3.23}$$

Es decir, podemos expresar la corriente de base como la contribución de dos diodos, con corriente inversa de saturación igual a I_S/β_F e I_S/β_R, respectivamente.

Evidentemente, ninguna de estas corrientes es igual a la de emisor ni a la de colector, por lo que debemos añadir más elementos al circuito equivalente. La corriente de emisor viene dada por:

$$I_E = \frac{I_{CC}}{\alpha_F} - I_{EC} = \left(\frac{\beta_F + 1}{\beta_F}\right) I_{CC} - I_{EC} = \\ = I_{CC} - I_{EC} + \frac{I_{CC}}{\beta_F} \tag{3.24}$$

Análogamente:

$$I_C = I_{CC} - I_{EC} + \frac{I_{EC}}{\beta_R} \tag{3.25}$$

El último sumando de las ecuaciones (3.24) y (3.25) ya está incluido en la expresión de la corriente de base, mientras que el resto es el mismo en ambas expresiones. Por tanto, podemos expresar estas ecuaciones mediante el siguiente circuito equivalente:

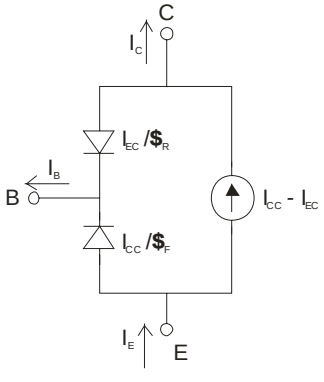


Figura 3.3.5 Versión de Spice del transistor bipolar

Modelos simplificados para análisis a mano

Según la región en la que se encuentre operando el transistor, se pueden eliminar algunos elementos de los anteriores modelos, simplificando notablemente el modelo. Además, los diodos se pueden

sustituir a su vez, para la resolución a mano de circuitos, por su correspondiente modelo (ideal con desplazamiento o lineal a trozos).

Activa

En la región activa podemos eliminar los elementos dependientes de la exponencial e^{V_{CB}/V_T}. El circuito basado en las ecuaciones de Ebers-Moll queda entonces como:

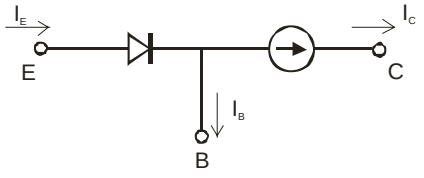


Figura 3.3.6 Versión simplificada del modelo de diodos acoplados válida en la región activa

El diodo se puede sustituir a su vez (para cálculos sencillos a mano) por una pila de 0.6-0.7 V en serie con una resistencia r_d.

No obstante, el circuito equivalente en activa usualmente empleado es el que se obtiene a partir de la versión de Spice:

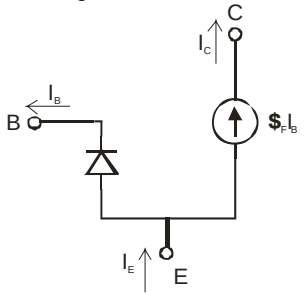


Figura 3.3.7 Modelo simplificado para el BJT en activa

Saturación

En saturación hay que usar el modelo completo. No obstante, se puede simplificar teniendo en cuenta que tenemos dos uniones polarizadas en directo. Por tanto, V_{BB} = 0.6 V y V_{BC} = 0.2 V.

Corte

En corte, sólo quedan en las ecuaciones de Ebers-Moll los términos correspondientes a las corrientes inversas de saturación de los diodos, que son muy pequeñas. Teniendo en cuenta la relación (3.20) podemos poner el circuito equivalente como:

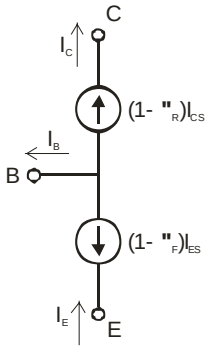


Figura 3.3.8 Modelo equivalente en la región de corte

Para un análisis sencillo a mano, basta con eliminar el transistor cuando éste se encuentra en la región de corte.

Ejercicio

Dibujar los circuitos equivalentes vistos en esta sección y en la anterior en el caso de que el transistor sea de tipo npn.

3.4 Características de transferencia. Polarización

Introducción

Ya hemos visto que un BJT puede trabajar en cuatro regiones de operación diferentes. Que lo haga en una u otra depende de las tensiones aplicadas a través de elementos externos (resistencias y baterías), que fijan el punto de trabajo del transistor (es decir, su polarización).

Los circuitos de polarización son los encargados de fijar el funcionamiento en continua (DC) del transistor, esto es, el valor de las tensiones aplicadas al BJT y de las corrientes que circulan por él.

En principio, hay seis variables para fijar: las tensiones entre los terminales (V_{BB}, V_{BC} y V_{EC}) y las corrientes que circulan (I_E, I_B e I_C). Sin embargo, sólo dos de ellas son independientes puesto que existen cuatro ecuaciones que relacionan estas seis variables: las dos ecuaciones de Ebers-Moll y las dos leyes de Kirchoff. Por tanto, basta con fijarnos en dos de estas variables e imponerlas externamente para determinar por completo el punto de polarización del transistor.

Vamos a ver ahora uno de los circuitos de polarización más sencillos. Además, introduciremos simultáneamente las características de transferencia (relaciones entre las variables, tensiones o corrientes, en el transistor) habitualmente empleadas, tanto para la descripción como para la polarización de un BJT.

Característica de transferencia I_B-V_{BE}

En la siguiente figura se muestra un circuito típico de polarización:

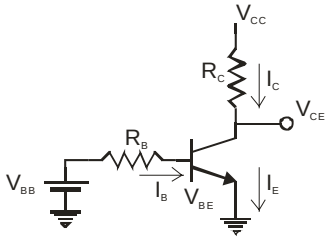


Figura 3.4.1 Circuito típico de polarización de un BJT

La malla formada por la pila V_{BB}, la resistencia de base (R_B) y la unión BE determinan la corriente de base (y, por tanto, la de colector y la de emisor). En efecto, como la corriente de base y la tensión de base-emisor son la solución del siguiente par de ecuaciones (en activa):

$$\begin{aligned} V_{BB} &= I_B R_B + V_{BE} \\ I_B &= \frac{I_S}{\beta_F}(e^{V_{BE}/V_T} - 1) \end{aligned} \qquad (3.26)$$

Obsérvese la completa analogía con la polarización de un diodo. La solución se puede obtener gráficamente mediante la intersección de la curva característica I_B-V_{BE} del transistor con la recta de carga impuesta por la resistencia R_B y la pila V_{BB} (como sucedía en el caso de un diodo).

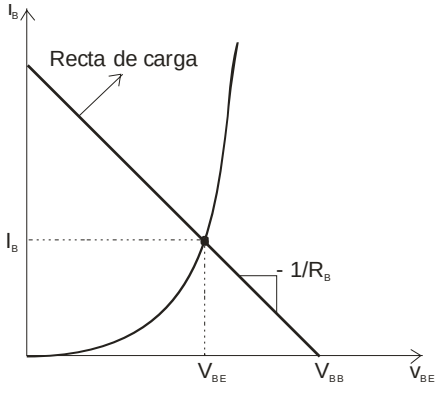


Figura 3.4.2 Curva característica i_B-v_{BE} de un transistor bipolar npn y recta de carga impuesta por la malla formada por la pila V_{BB}, la resistencia de base R_B y la unión BE en el circuito de polarización de la Figura 3.4.1.

Normalmente, para realizar un análisis y diseño a mano de un circuito de polarización no se suele resolver el anterior sistema de ecuaciones, sino que se da un valor fijo a la tensión V_{BB} (aproximadamente 0.6-0.7 V) y se calcula la corriente de base como:

$$I_B = \frac{V_{BB} - V_{BE}}{R_B} \tag{3.27}$$

Característica de transferencia I_C-V_{CE}

Estas características de transferencia tienen la forma mostrada en la siguiente figura:

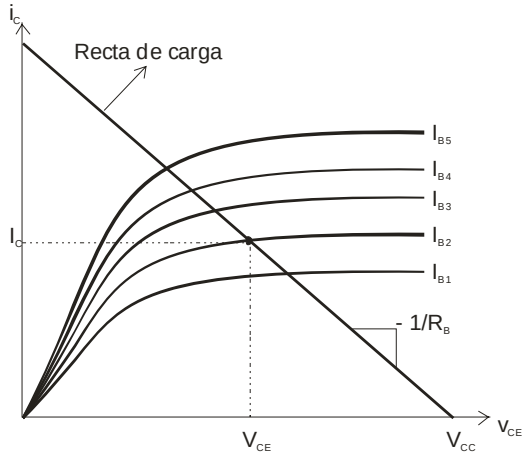


Figura 3.4.3 Familia de curvas características i_c-v_{ce} con i_b como parámetro

Cada una de las curvas se corresponde con un valor constante de la intensidad de base (I_B) o, lo que es lo mismo, de la tensión V_{BE}. Para entender por qué tiene esta forma la característica $I_C - V_{CE}$ recordemos las ecuaciones de Ebers-Moll:

$$\begin{aligned} I_E &= \frac{I_S}{\alpha_F}\left(e^{V_{EB}/V_T} - 1\right) - I_S\left(e^{V_{CB}/V_T} - 1\right) \\ I_C &= I_S\left(e^{V_{EB}/V_T} - 1\right) - \frac{I_S}{\alpha_R}\left(e^{V_{CB}/V_T} - 1\right) \end{aligned} \tag{3.28}$$

En la región activa:

$$I_E = \frac{I_S}{\alpha_F} e^{V_{EB}/V_T}$$
$$I_C = I_S e^{V_{EB}/V_T} \qquad (3.29)$$
$$I_B = \frac{I_S}{\beta_F} e^{V_{EB}/V_T}$$

Como vemos, fijado el valor de V_{BB} está fijado también el valor de las corrientes que circulan, que son independientes de la tensión en la otra unión (V_{CB}) y, por tanto, también de la tensión $V_{CE} = V_{CB} + V_{BE}$. Por eso, la curva I_C-V_{CE} es casi plana (tiene una pequeña pendiente debida al Efecto Early, que comentaremos posteriormente).

Fijado el valor de *V*BB, al disminuir el valor de V_{CB} va cobrando importancia el otro sumando de las ecuaciones de Ebers-Moll, y va disminuyendo el valor de las corrientes de emisor y de colector. Cuando V_{CB} es igual a cero, estamos en la frontera entre las regiones activa y saturación. Para valores inferiores, la unión CB está en directo y la tensión V_{CB} es pequeña. Las regiones activa y saturación están separadas en la gráfica I_C-V_{CE} por una curva de tipo exponencial. En efecto:

$$V_{CB} = 0 \Rightarrow V_{CE} = V_{CB} \Rightarrow I_C = I_S e^{V_{EB}/V_T}$$

3.5 Comportamiento dinámico

Introducción

Hasta ahora se ha estudiado la relación entre las corrientes y las tensiones en el BJT cuando éstas son estacionarias, no cambian con el tiempo. Sin embargo, en la operación normal del BJT pasaremos de un estado a otro o le aplicaremos señales variables con el tiempo, por lo que a continuación se va a estudiar cómo responde el transistor bipolar en estas situaciones.

BJT en conmutación

Introducción

En aplicaciones digitales, interesa que los BJTs actúen como interruptores. Un interruptor ideal cumple con las siguientes especificaciones:

- Estado ON: entre sus extremos caen 0 V para cualquier intensidad circulando entre sus extremos
- Estado OFF: bloquean el paso de corriente (I = 0 A) para cualquier tensión aplicada entre sus extremos

Un BJT en los estados de saturación y corte cumple aproximadamente con estas especificaciones, puesto que:

- En saturación, $V_{CE} = 0.2$ (es decir, casi 0 V) independientemente de la corriente de colector.
- En corte, I es prácticamente igual a 0 A, independientemente de las tensiones aplicadas (siempre que mantengan las uniones en inversa)

Antes de analizar el transitorio de conmutación entre estos estados, vamos a analizar con más detalles estas regiones de operación.

BJT en corte

Como ya explicó anteriormente, las corrientes en el estado de corte son muy pequeñas y pueden despreciarse en la mayoría d los casos. En los límites de la base con las zonas de carga espacial, tenemos un defecto de minoritarios (las uniones están en inverso). En un transistor pnp:

$$\Delta p_E = -p_{n0} = \Delta p_C \tag{3.30}$$

Puesto que la corriente es despreciable, podemos suponer que en toda la base el gradiente de concentración es nulo y, por tanto, la concentración de minoritarios es constante (y nula). Es decir, el exceso de minoritarios viene dado por la siguiente expresión (ver también Figura 3.5.1):

$$\delta p_n(x_n) = -p_{n0} \tag{3.31}$$

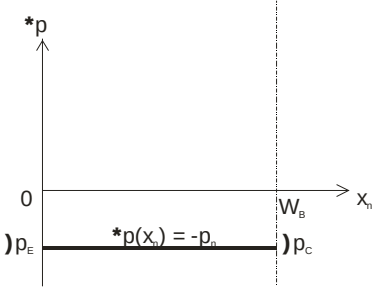


Figura 3.5.1 Exceso de minoritarios en la base cuando el transistor se encuentra en corte

Conmutación

Hacer pasar el BJT de corte a saturación y viceversa requiere un cierto tiempo, no es una transición instantánea. El tiempo necesario se emplea en eliminar o aportar la carga en la base.

La Figura 3.5.2 ilustra cómo cambia la carga en la base y la corriente en el colector durante una conmutación de corte a saturación seguida de otra en sentido contrario. Es importante señalar que mientras que el transistor esté en saturación, no cambia la corriente de colector.

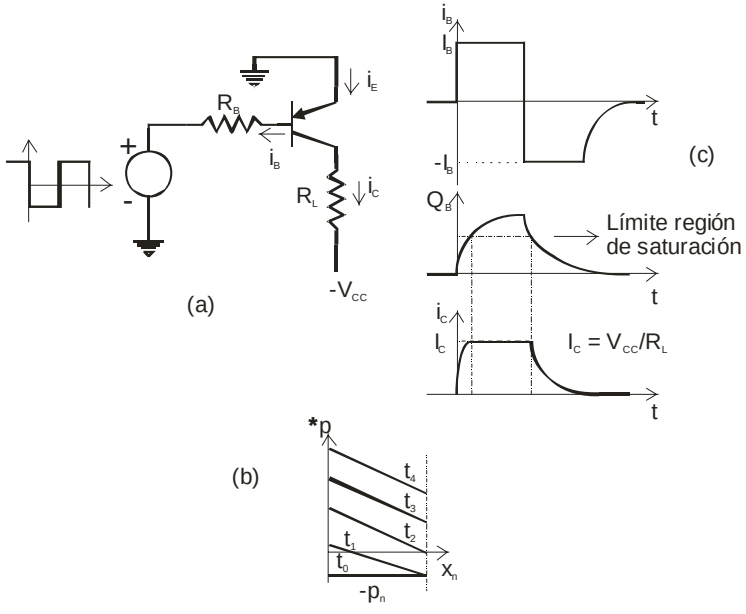


Figura 3.5.2 Transitorios de conmutación de corte a saturación y de saturación a corte. Circuito (a), evolución del exceso de minoritarios en la base (b), de la carga total en la base y de la corriente de colector (c). La carga en la base en el instante t2 corresponde con el límite de la región de saturación.

Modelo de pequeña señal

En la Figura 3.5.3 se ilustra el modelo en π de pequeña señal de un transistor bipolar npn.en activa.

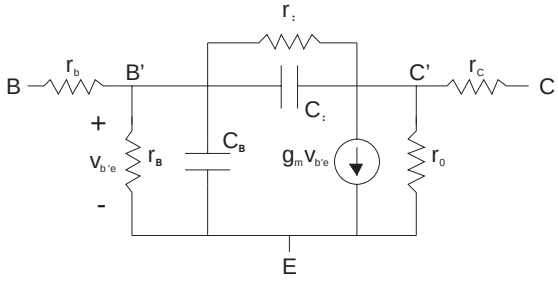


Figura 3.5.3 Modelo de pequeña señal de un transistor bipolar npn

Donde:

- $r_{bb'}$ y $r_{cc'}$ dan cuenta de la caída de tensión desde los terminales del dispositivo hasta la región en donde está el BJT
- $r_{ee'}$ es despreciable porque el terminal de emisor está más próximo a la conexión correspondiente y porque el dopado de emisor es más alto (la resistencia es menor)
- Las variaciones en la corriente de emisor (i_e) debidas a un cambio en la tensión en la unión BE ($v_{be'}$) las descomponemos en las variaciones en la corriente de colector (i_c) más las variaciones en la corriente de base (i_b):

$$i_e = i_c + i_b$$

- $g_m v_{be'}$ reproduce la variación en la corriente de colector causada por un cambio ($v_{be'}$) en la tensión en la unión BE. Por tanto, g_m viene dada por:

$$g_m = \left(\frac{\partial i_C}{\partial v_{BE}} \right)_{V_{CE}=cte} = \frac{I_C}{V_T} \tag{3.32}$$

- r_π da cuenta de la variación en la corriente de base:

$$r_\pi^{-1} = \frac{\partial i_B}{\partial v_{BE}} = \frac{I_B}{V_T} = \frac{I_C}{\beta_F V_T} \tag{3.33}$$

- Las capacidades C_π y C_μ están asociadas a las uniones BE y BC, respectivamente:

$$C_\pi = C_d + C_{je} = \tau_F \frac{I_C}{V_T} + \frac{CJE}{\left(1 - \frac{V_{BE}}{V_i}\right)^m} \approx \tau_F g_m + 2CJE \tag{3.34}$$

$$C_\mu = C_{cj} = \frac{CJC}{\left(1 - \frac{V_{BC}}{V_i}\right)^m} \tag{3.35}$$

- Finalmente, r_0 da cuenta del efecto Early:

$$r_0^{-1} = \left(\frac{\partial i_C}{\partial v_{CE}} \right)_{V_{BE}=cte} = \frac{I_C}{V_A} \tag{3.36}$$

3.6 Efectos de segundo orden

A continuación se comentará brevemente algunos efectos no considerados en el tratamiento simplificado que condujo a las ecuaciones de Ebers-Moll.

Deriva en la base

La diferente concentración de impurezas en la base da lugar a la presencia de un campo eléctrico que altera la difusión de minoritarios en la base.

A veces este efecto se provoca, para conseguir que el campo eléctrico generado esté a favor del movimiento de los minoritarios, disminuyendo su tiempo de tránsito por la base y aumentando, por tanto, α_F. En este caso, se obtienen transistores bipolares de deriva.

Estrechamiento de la base

Según el modelo de Ebers-Moll, en activa la corriente de emisor es independiente de la tensión V_{CE} entre colector y emisor o, lo que es lo mismo, es independiente de la tensión inversa aplicada en la unión BC. Sin embargo, cuanto mayor sea esta tensión, mayor es la zona de carga espacial y menor es la anchura efectiva de la base. Esto hace que la probabilidad de recombinarse en la base sea menor y aumenta, por tanto, el factor de transporte α_T y, consiguientemente, α_F y β_F.

Este fenómeno (conocido como efecto Early) puede modelarse incluyendo un factor adicional en la expresión que nos proporciona la corriente de colector en activa:

$$I_C = I_s e^{V_{BE}/V_T}\left(1+\frac{V_{CE}}{V_A}\right), \tag{3.37}$$

donde a V_A se le conoce como tensión Early.

Ruptura por avalancha

En activa la unión BC está en inverso y, al igual que sucede con una unión PN, pueden producirse mecanismos de ruptura por avalancha que incrementen bruscamente la corriente de colector. La tensión de ruptura depende de la configuración y se comprueba que la correspondiente a emisor común (BV_{CE0}) es menor que la base común (BV_{CB0}).

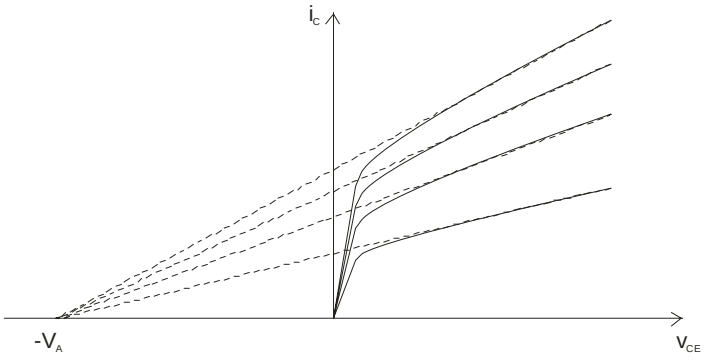


Figura 3.6.1 Efecto de la modulación de la anchura de la base y tensión Early

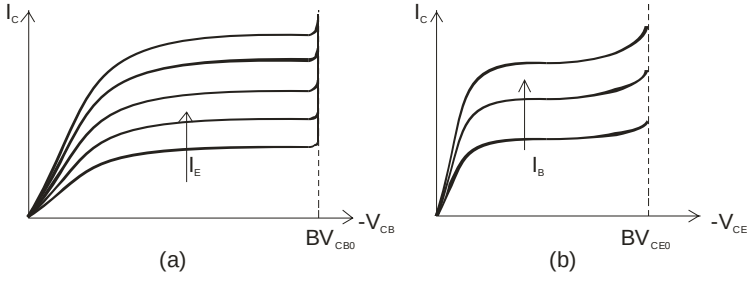


Figura 3.6.2 Ruptura por avalancha en la unión BC con una configuración de base común (a) y de emisor común (b)

Dependencia de β_F con la corriente de colector

La ganancia de corriente β_F resulta ser dependiente de la tensión de colector aplicada, siendo menor en las zonas de baja y alta corriente que en la zona central de corriente intermedia.

A bajas corrientes, la corriente de colector disminuye debido a que la fracción de portadores inyectados por el emisor que se recombinan en la zona de carga espacial antes de llegar a la base (y, por tanto, al colector) no es despreciable. Con altos niveles de inyección, el efecto Kirk provoca la disminución de la corriente de colector. Del mismo modo que el efecto Early provoca el aumento de la corriente de colector debido a una disminución de la anchura efectiva de la base, el efecto Kirk provoca la disminución de la corriente de colector al aumentar el ancho efectivo de la base.

Este aumento del ancho efectivo de la base está provocado por el hecho de tener una gran concentración de portadores en tránsito por la zona de carga espacial (por ejemplo, huecos en un transistor pnp). Esto hace que tengamos una densidad de carga positiva adicional, que se suma a la de impurezas ionizadas positivamente en la z.c.e. correspondiente a la base y que resta a la densidad de cargas negativas debida a las impurezas

ionizadas en la z.c.e. del colector. Por tanto, para tener la misma densidad de carga que (con baja inyección) se corresponde con la caida de tensión V_{CB}, el ancho de la z.c.e en la base tiene que ser menor y en el colector mayor.

REFERENCIAS

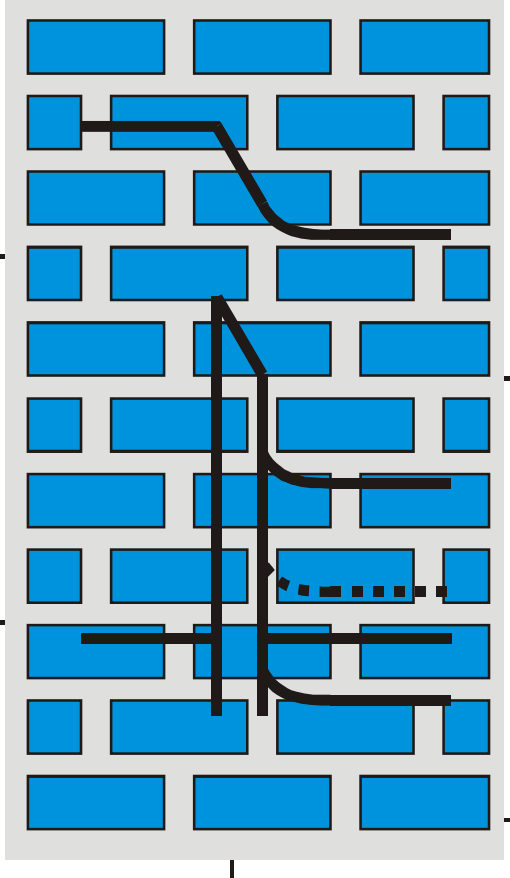
[1] B.G. Streetman, S. Banerjee, "Solid State Electronic Devices", Prentice Hall, 2000.
[2] S. Dimitrijev, "Understanding Semiconductor Devices", Oxford University Press, 2000.

4

Capítulo

La Estructura Metal Aislante Semiconductor

Diagrama de bandas de una estructura MIS en equilibrio



Índice

Objetivos

- Definir el concepto de transistor de efecto campo.
- Clasificar distintos tipos de transistores de efecto campo.
- Presentar el dispositivo de efecto campo más usado hoy en día en electrónica: el MOSFET.
- Analizar de forma separada, como dispositivo aparte, una parte esencial de su estructura: el dispositivo metal-aislante-semiconductor (MIS).
- En esta estructura el objetivo prioritario será establecer una relación entre la tensión aplicada a los terminales externos con magnitudes eléctricas internas como son la caída de potencial en el óxido y en el semiconductor, y las cargas existentes en el semiconductor, en el aislante y en la interfaz silicio-aislante.

Palabras Clave

Transistor de efecto campo.
Resistencia controlada por tensión.
Corriente controlada por tensión.
Dispositivos unipolares.
Terminales de puerta, fuente y drenador.
Estructura metal-aislante-semiconductor.
Función trabajo de un material.
MIS en acumulación.
Región de vaciamiento "depletion".
MIS en inversión.
Tensión de banda plana.
Potencial de superficie.
Carga en el óxido.
Estados superficiales.
Carga en el semiconductor.
Capacidad MIS.

4.1 Transistores de efecto campo

Transistor de efecto campo
Es un dispositivo de tres terminales, unipolar. La corriente que circula entre dos de ellos se controla por la tensión aplicada al tercero. Dicho control se lleva a cabo modificando el espesor del canal por el que circulan las cargas.

Son dispositivos de tres terminales unipolares pues en su funcionamiento interviene solo un tipo de portadores de carga. A través de uno de esos terminales, denominado puerta, aplicaremos un campo eléctrico, con el cual controlaremos la corriente que circula entre los otros dos terminales del dispositivo. A estos últimos los denominaremos fuente (donde salen los portadores de carga) y drenador (donde se recogen los portadores).

La idea original de transistor de efecto campo la encontramos en unas patentes de Lilienfeld de 1926-28, donde se presentaba una estructura en la que se podía controlar la carga de un canal semiconductor mediante una lámina metálica 'G'.

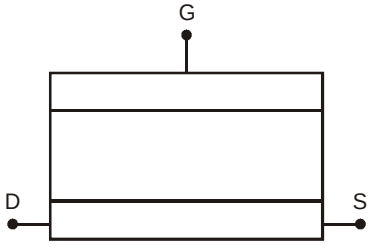


Figura 4.1.1 Idea original de transistor de efecto campo: una lámina metálica 'G' controla la carga de un semiconductor (entre los terminales D y S).

El desarrollo de este tipo de transistores fue posterior al de los transistores bipolares. Recordemos que tres investigadores, Shockley, Brattain y Bardeen recibieron el premio Nobel por sus trabajos encaminados a la invención del transistor bipolar de unión a finales de la década de los 40. Sin embargo, hoy en día el número de transistores de efecto campo que se producen en un año, en particular los transistores de efecto campo metal óxido semiconductor (MOSFET), superan con creces a los transistores bipolares.

El estudio de los transistores de efecto campo se va a iniciar con el MOSFET (Figura 4.1.2) por ser el transistor más empleado en electrónica [1]. En esta figura también se muestran otros transistores de efecto campo como el transistor de efecto campo de unión (JFET). La corriente que circula entre fuente y drenador se controla modificando la tensión aplicada a una unión en inverso. Sus aplicaciones principales se limitan a lo que se conoce como "front-end electronics", que podría traducirse como electrónica de choque. Existen ciertas situaciones donde las condiciones de operación son extremadamente desfavorables, como existencia de radiación. El transistor más resistente es el JFET, eligiéndose para trabajar en ellas. El análisis de este transistor se llevará a cabo en el capítulo 6. En el capítulo 7 se describirá el funcionamiento del transistor de efecto campo metal semiconductor (MESFET). La diferencia

con el anterior es que emplea una unión entre un metal y un semiconductor en lugar de una unión semiconductora como parte esencial de la estructura. Una variante de este transistor es el transistor de efecto campo de alta movilidad (HFET), con aplicaciones en la región de microondas, y que emplea heterouniones semiconductoras para conseguir un canal de electrones de movilidad elevada. Todas estas estructuras se pueden contemplar en la Figura 4.1.2.

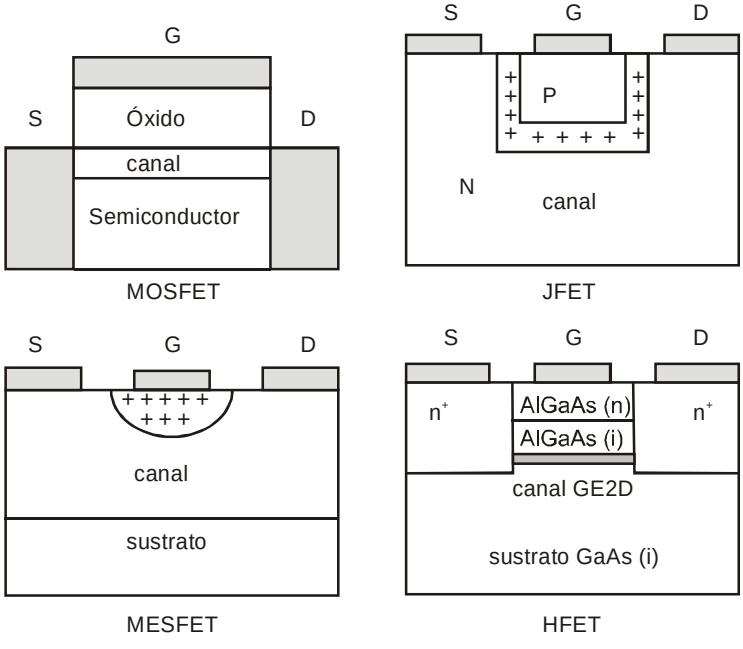


Figura 4.1.2 Diferentes estructuras de transistor de efecto campo: MOSFET, JFET, MESFET Y HFET.

4.2 Estructura MIS

Antes de analizar la estructura completa de un MOSFET la estudiaremos paso a paso. En primer lugar se estudiará un dispositivo de dos terminales conocido con el nombre de metal aislante semiconductor (MIS), que realmente es el pilar en el que se sustenta dicha estructura. Se empleará la palabra aislante para hacer más general este dispositivo, aunque realmente la mayor parte de los casos el aislante será el SiO_2.

Consideremos las tres partes que constituyen la estructura MIS por separado (Figura 4.2.1). Las vamos a caracterizar por su función trabajo, ϕ, (ϕ_s si se trata del semiconductor y ϕ_m si se trata del metal).

Función trabajo, ϕ: Diferencia entre el nivel del vacío, E_0, (nivel que tendría un electrón libre en el vacío) y el nivel de Fermi del material, E_F.

En la Figura 4.2.1 también se representa la afinidad electrónica del semiconductor, χ_s. Vamos a considerar que el semiconductor es tipo P con lo que el nivel de Fermi está situado próximo a la banda de valencia. Por otro lado consideremos el caso en el que la función trabajo del semiconductor es menor que la del metal: $\phi_s < \phi_m$.

Estructura MIS en equilibrio

Si encontráramos la forma de que estos tres materiales entraran en contacto y los dejáramos evolucionar hacia el equilibrio, la estructura de bandas de la nueva estructura sería diferente a la que presentan los tres materiales por separado. Para que se alcance el equilibrio, y el nivel de Fermi sea el mismo en el metal y en el semiconductor, necesariamente debe haber una redistribución de carga. En la Figura 4.2.1 vemos que el nivel de Fermi del semiconductor, E_{Fs}, está por encima del nivel de Fermi del metal, E_{Fm},. Para que se igualen los niveles de Fermi, en el semiconductor se debería desocupar la banda de valencia de electrones y en el metal debería ocuparse la banda de conducción con más electrones.

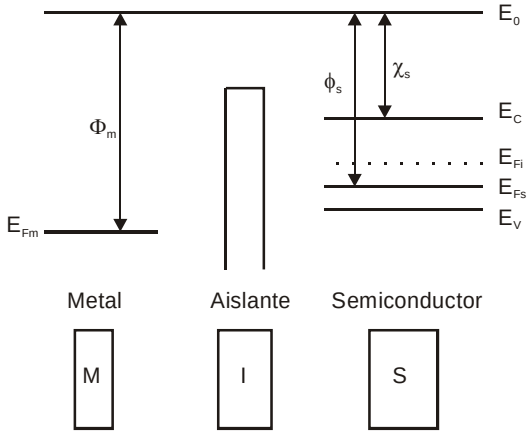


Figura 4.2.1 Diagrama de bandas de cada una de las partes constituyentes de la estructura MIS analizadas por separado.

Al desocuparse la banda de valencia en el semiconductor, es decir al llenarse con más huecos, la banda de valencia tiende a acercarse al nivel de Fermi (Figura 4.2.2). El acercamiento es mayor cuanto más cerca nos encontremos del aislante. Lejos del aislante, la acción del contacto deja de sentirse en el semiconductor con lo que la banda de valencia está situada con respecto al nivel de Fermi en la misma posición que en el caso de encontrarse el semiconductor aislado. La banda de conducción se curva exactamente igual que la de valencia pues el ancho de la banda prohibida es una constante del semiconductor.

Región de acumulación en un MIS:
Los portadores mayoritarios del semiconductor se acercan a la superficie próxima al aislante, aumentando su concentración en esa región.

En esta situación se dice que la estructura se encuentra en acumulación, por los huecos acumulados cerca de la superficie con el aislante.

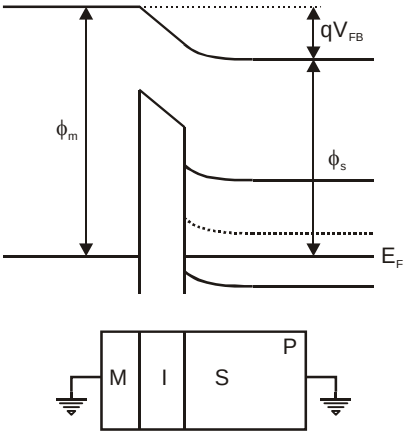


Figura 4.2.2 Diagrama de bandas de la estructura MIS en equilibrio.

4.3 Estructura MIS polarizada

Recordemos que nuestro objetivo con cualquier dispositivo es aplicarle una diferencia de potencial entre sus terminales y encontrar la corriente que circula por los mismos. En esta estructura la conexión de una fuente de alimentación entre el metal y el semiconductor no va a dar como resultado una corriente pues entre ambos se encuentra un aislante que actúa de barrera. Sin embargo, es de gran interés analizar qué ocurre en la estructura, pues será fundamental para estudiar el MOSFET. Definamos en primer lugar sus terminales (Figura 4.3.1). Llamaremos puerta al terminal conectado al metal y sustrato al terminal conectado al semiconductor

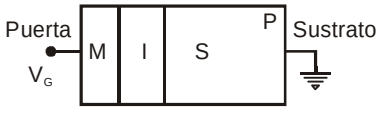


Figura 4.3.1 Estructura MIS polarizada.

Tensión de banda plana

Cuando aplicamos una diferencia de potencial entre el metal y el semiconductor, aunque no exista una corriente, sí habrá una reorganización de la carga en el mismo. Lo que es evidente es que si aplicamos una tensión positiva a la puerta alejaremos los huecos de la superficie con el aislante. Si la tensión es negativa lo que haremos será atraer más hacia esa superficie. Lo primero que nos vamos a preguntar es qué tensión debemos aplicarle al metal para que las bandas sean planas, es decir para conseguir la misma concentración de huecos en todo el semiconductor e igual a la que tendría en el caso de estar aislado (Figura

4.3.2). A esa tensión la llamaremos tensión de banda plana, V_{FB}. Para lograr esa situación debemos alejar los huecos acumulados en superficie. Habrá que aplicar una tensión positiva al metal. El nivel de Fermi en el metal quedará por tanto desplazado del nivel de Fermi en el semiconductor en una cantidad $-qV_{FB}$. En el caso de que no existan cargas en la estructura (cargas en el óxido o estados superficiales) que den lugar a un campo eléctrico adicional la tensión de banda plana será igual a la diferencia de funciones trabajo entre el metal y el semiconductor

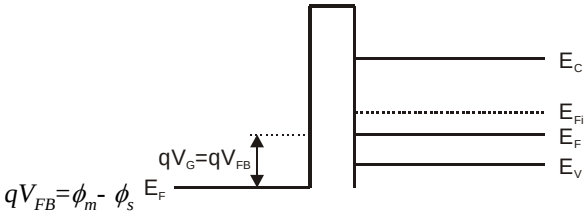


Figura 4.3.2 Diagrama de bandas de la estructura MIS para una tensión de puerta igual a la tensión de banda plana.

Región de vaciamiento o "deplexión"

Región de deplexión:
Los portadores mayoritarios del semiconductor se alejan de la superficie próxima al aislante dejando en esa región una carga fija formada por impurezas ionizadas.

Cuando aplicamos una tensión positiva al metal mayor que la tensión de banda plana se siguen alejando huecos de la superficie, creándose un zona de carga espacial, compuesta por impurezas ionizadas negativamente (Figura 4.3.3). Las bandas en el semiconductor se curvan de forma que reflejen este fenómeno. Cerca de la superficie la banda de valencia se aleja del nivel de Fermi indicando la disminución de huecos en esa región. A la curvatura total de las bandas la llamaremos $q\psi_s$, donde ψ_s se denomina potencial de superficie.

Potencial de superficie, ψ_s:
Referido a la curvatura máxima de las bandas en el semiconductor corresponde al valor del potencial en la superficie del semiconductor con el aislante.

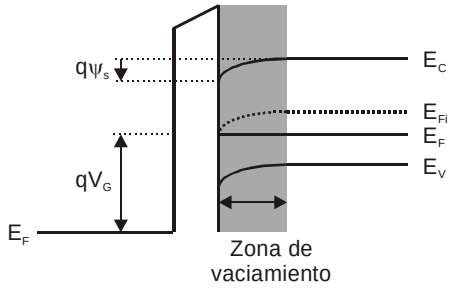


Figura 4.3.3 Diagrama de bandas de la estructura MIS en deplexión. La región sombreada indica la zona del semiconductor donde se han eliminado los huecos por la aplicación de una tensión positiva a la puerta.

Región de inversión

Si se aumenta la tensión positiva aplicada al metal no sólo se seguirán repeliendo huecos de la superficie sino que se atraerán portadores minoritarios (electrones). En la superficie se creará una capa

de inversión (carga móvil) que convivirá con la zona de carga espacial (carga fija) situada a continuación de la capa de inversión (Figura 4.3.4). Es primordial conocer el límite entre la región de vaciamiento y la región de inversión, es decir cuándo existe solo carga fija y cuándo aparece la carga móvil. Para contestar esta pregunta debemos fijarnos en el diagrama de bandas de la estructura. Al aumentar la tensión aplicada al metal la banda de valencia se sigue alejando del nivel de Fermi, pero simultáneamente la banda de conducción se acerca al nivel de Fermi. Llegará un momento en el que la diferencia entre el fondo de la banda de conducción y el nivel de Fermi en la superficie se aproxime a la separación entre el nivel de Fermi y el máximo de la banda de valencia en la zona neutra del semiconductor, lejos de la superficie. Ese será el momento en el que se defina el límite entre las regiones de deplexión e inversión. Se define dicho límite entre regiones cuando se cumpla $\psi_s = 2\phi_F$ con $\phi_F = (E_{Fi}(\infty) - E_F)/q$.

Región de inversión:
Los portadores minoritarios del semiconductor se acercan a la superficie próxima al aislante. Se produce un cambio en el tipo de portadores existentes en esa región.

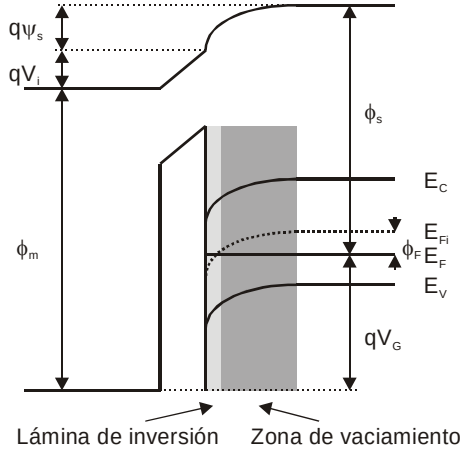


Figura 4.3.4 Diagrama de bandas de la estructura MIS en inversión. La región sombreada más cercana al aislante corresponde a la zona del semiconductor donde han aparecido electrones.

La condición anterior no es suficiente para determinar la separación entre las regiones de vaciamiento e inversión. Lo que realmente se necesita conocer es la tensión aplicada a la puerta para que aparezca la lámina de inversión de carga. A partir del diagrama de bandas de la Figura 4.3.5 y haciendo uso de la ley de Gauss se puede relacionar el potencial de superficie con la tensión entre terminales V_G.

Para ello se definen dos volúmenes delimitados por sendas superficies cerradas, uno dentro del aislante y otro en el semiconductor, tal y como se ven en la Figura 4.3.5. El volumen localizado en el semiconductor comprende toda la carga por unidad de superficie existente en este material, Q_S. Uno de los extremos de ese volumen se coloca dentro del semiconductor pero muy próximo al aislante. En esa cara el campo

eléctrico se denomina E_S. El otro extremo se toma dentro de la zona neutra del semiconductor donde el campo eléctrico es nulo. La otra superficie de Gauss se sitúa dentro del aislante de manera que incluya las posibles cargas fijas por unidad de superficie que existen en el óxido, Q_{ox}, y la densidad de estados superficiales de la interfaz semiconductor-aislante, Q_{SS}. Una superficie de contorno se elige dentro del aislante de manera que el campo sea igual al campo del óxido, E_i, y la otra en la interfaz semiconductor-aislante pero dentro del semiconductor, de manera que el campo corresponda al del semiconductor en ese punto, E_S, y el volumen contenga los estados superficiales.

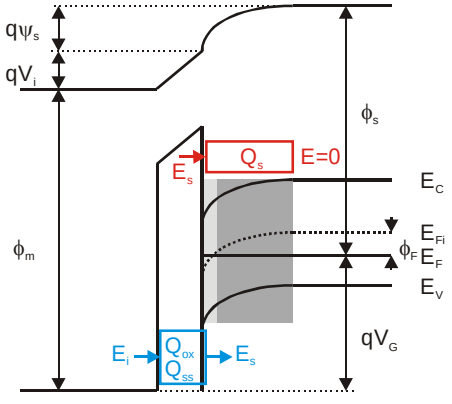


Figura 4.3.5 Relación entre el campo y las cargas existentes en diferentes regiones de la estructura MIS.

Aplicando la ley de Gauss a las dos superficies se obtiene:

$$Q_s = -\varepsilon_s E_s$$
$$Q_{ox} + Q_{ss} = -\varepsilon_i E_i + \varepsilon_s E_s = -\varepsilon_i \frac{V_i}{d_i} - Q_s \qquad (4.1)$$

donde ε_s y ε_i son las constantes dieléctricas del semiconductor y el aislante respectivamente, V_i es la tensión que cae en el aislante y d_i es el espesor del aislante.

Analizando el diagrama de bandas de la Figura 4.3.5 se puede relacionar la tensión externa V_G con el potencial de superficie ψ_s:

$$qV_G = \phi_m - \phi_s + q\psi_s + qV_i \,. \qquad (4.2)$$

Introduciendo las ecuaciones (4.1) en esta última se obtiene:

$$qV_G = \phi_{ms} + q\psi_s - q\frac{Q_s + Q_{ss} + Q_{ox}}{C_{ox}} \,, \qquad (4.3)$$

donde C_{OX} es la capacidad del aislante por unidad de superficie, $C_{ox} = \varepsilon_i / d_i$ y d_i es el espesor del aislante.

Si se quiere calcular ahora la tensión de banda plana bastaría con anular la carga del semiconductor y el potencial de superficie. La tensión

de banda plana queda corregida por las cargas en el óxido y en la interfaz Si-SiO_2:

$$qV_{FB} = \phi_{ms} - q\frac{Q_{ss} + Q_{ox}}{C_{ox}}. \qquad (4.4)$$

La ecuación (4.3) se puede expresar por tanto:

$$qV_G = qV_{FB} + q\psi_s - q\frac{Q_s}{C_{ox}}. \qquad (4.5)$$

4.4 Determinación de la carga en el semiconductor

La ecuación (4.5) no es definitiva por cuanto se desconoce cuál es el valor de la carga que existe en el semiconductor para un valor dado de la tensión de puerta, $Q_S = Q_S(V_G)$. Ni siquiera se conoce la relación carga del semiconductor con el potencial de superficie, $Q_S = Q_S(\psi_S)$. Esta relación se puede obtener resolviendo la ecuación de Poisson en el semiconductor:

$$\frac{d^2V}{dx^2} = -\frac{\rho(x)}{\varepsilon_s} = -\frac{q(p - n - N_A)}{\varepsilon_s}$$
$$p = p_{po}e^{-\frac{qV}{KT}},\ n = n_{po}e^{\frac{qV}{KT}}, \qquad (4.6)$$
$$N_A = p_{po} - n_{po},$$

donde p es la concentración de huecos, n la de electrones, N_A la concentración de impurezas aceptadoras, p_{p0} es la concentración de huecos en equilibrio, n_{p0} la concentración de electrones en equilibrio, ε_S es la constante dieléctrica del semiconductor y $V = V(x)$ es el potencial en un punto x del semiconductor (Figura 4.4.1). Se ha tomado como origen $x = 0$ la interfaz aislante-semiconductor, y como origen de potenciales el potencial de los electrones de la banda de conducción situados en la zona neutra del semiconductor.

Introduciendo los valores de n, p y N_A en la ecuación de Poisson (4.6) y relacionando el campo eléctrico con el potencial se obtiene:

$$\frac{d^2V}{dx^2} = \frac{q}{\varepsilon_s}\left[n_{po}(e^{\frac{qV}{KT}} - 1) - p_{po}(e^{-\frac{qV}{KT}} - 1)\right] = -\frac{dE}{dx} = E\frac{dE}{dV}. \qquad (4.7)$$

Integrando entre el borde de la zona de carga espacial con la zona neutra, donde el campo y el potencial son nulos, y un punto x del semiconductor se llega a:

$$E^2 = 2\frac{q}{\varepsilon_s}\int_0^V \left(n_{po}(e^{\frac{q'V}{KT}}-1) - p_{po}(e^{\frac{q'V}{KT}}-1)\right)d'V$$
$$= 2\frac{KT}{\varepsilon_s}\left(p_{po}(e^{\frac{qV}{KT}}+\frac{qV}{KT}-1)+n_{po}(e^{\frac{qV}{KT}}-\frac{qV}{KT}-1)\right) \quad (4.8)$$

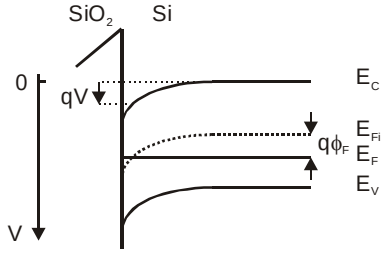


Figura 4.4.1 Diagrama de bandas en el semiconductor de la estructura MIS. Definición del potencial $V = V(x)$ en un punto x.

Introduciendo la longitud de Debye, L_{Dp}, y la función *f* definidas como:

$$L_{Dp} = \sqrt{\frac{\varepsilon_s KT}{q^2 N_A}},$$
$$f = \pm\left((e^{\frac{qV}{KT}}+\frac{qV}{KT}-1)+\frac{n_{po}}{p_{po}}(e^{\frac{qV}{KT}}-\frac{qV}{KT}-1)\right)^{\frac{1}{2}}, \quad (4.9)$$

donde el signo negativo se emplea para semiconductores N y el signo positivo para semiconductores P, se puede escribir el campo eléctrico como:

$$E = \sqrt{2}\frac{KT}{qL_{Dp}}f\ . \quad (4.10)$$

Se puede calcular el campo en la superficie del semiconductor, E_S, sin más que particularizar el valor del potencial *V* por el potencial de superficie ψ_s:

$$E_s = \sqrt{2}\frac{KT}{qL_{Dp}}f\left(\frac{q\psi_s}{KT}\right). \quad (4.11)$$

Conocido este campo podemos calcular la carga que existe en el semiconductor.

$$|Q_s| = \varepsilon_s |E_s|. \quad (4.12)$$

En acumulación e inversión los términos que dominan en la función *f* son los exponenciales, obteniéndose un crecimiento de la carga

en estas regiones a medida que aumenta el valor absoluto de la tensión aplicada a la puerta:

$$|Q_s| \propto e^{\frac{q|\psi_s|}{2KT}}. \tag{4.13}$$

En la región de vaciamiento quienes dominan son los términos lineales dentro de la raíz, pudiéndose aproximar la carga por:

$$|Q_s| \approx \varepsilon_s \frac{\sqrt{2}KT}{qL_{Dp}} \sqrt{\frac{q\psi_s}{KT} - 1} \approx \varepsilon_s \frac{\sqrt{2}KT}{qL_{Dp}} \sqrt{\frac{q\psi_s}{KT}} = \sqrt{2\varepsilon_s q N_A \psi_s}. \tag{4.14}$$

A partir de esta expresión se puede estimar el ancho de la zona de carga espacial sin más que igualar:

$$|Q_S| = qN_A W = \sqrt{2\varepsilon_s q N_A \psi_s}, \tag{4.15}$$

de donde se obtiene una expresión muy similar a la de la anchura de la zona de carga espacial en una unión pn:

$$W = \sqrt{\frac{2\varepsilon_s \psi_s}{q N_A}}. \tag{4.16}$$

4.5 Capacidad de la estructura MIS

Si se somete a la estructura a variaciones de tensión la carga almacenada en el semiconductor también se verá alterada. Esto da lugar a unos efectos capacitivos que se pueden cuantificar sin más que derivar (4.12) respecto al potencial de superficie:

$$C_s = A\frac{dQ_s}{d\psi_s},$$

$$C_s = A\left(\frac{\varepsilon_s q^2 N_A}{2KT}\right) \frac{\left|(1 - e^{\frac{-q\psi_s}{KT}}) + \frac{n_{po}}{p_{po}}(e^{\frac{q\psi_s}{KT}} - 1)\right|}{\left[(e^{\frac{-q\psi_s}{KT}} + \frac{q\psi_s}{KT} - 1) + \frac{n_{po}}{p_{po}}(e^{\frac{q\psi_s}{KT}} - \frac{q\psi_s}{KT} - 1)\right]^{\frac{1}{2}}} \tag{4.17}$$

Si se representa esta función se puede ver que el valor de la capacidad tiende rápidamente a infinito cuando la estructura se adentra mucho en acumulación o inversión. Esto es lógico pues a medida que se penetra en esas dos regiones la concentración de portadores móviles irá en aumento, incrementándose también la capacidad.

Sin embargo, una cosa es el valor teórico de la capacidad asociada al semiconductor y otra bien distinta es la capacidad real de la estructura. Ésta es la combinación serie de la capacidad que se acaba de calcular con la capacidad que presenta el óxido, C_{OX} tal y como se muestra en la Figura 4.5.1.

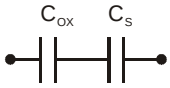


Figura 4.5.1 Modelo de capacidad de la estructura MIS. Está compuesta por la capacidad del aislante y la capacidad del semiconductor $C = C_{OX} \,//\, C_S$.

Como la capacidad del óxido toma un valor constante, independientemente de la tensión externa aplicada, la capacidad equivalente en las regiones de inversión y acumulación coincidirá con la capacidad del óxido C_{OX}. (Figura 4.5.2)

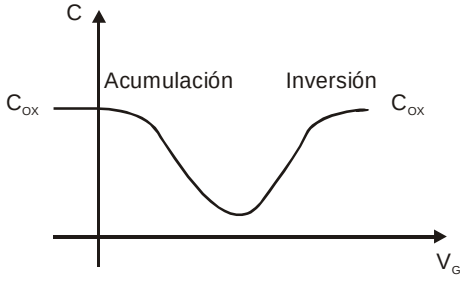


Figura 4.5.2 Capacidad vista entre los terminales de la estructura MIS.

Esta curva es válida a bajas frecuencias. La razón es que los portadores que constituyen el canal son portadores minoritarios que provienen de todo el semiconductor. Si aumentamos la frecuencia puede ocurrir que no le demos tiempo a estos portadores minoritarios a acercarse y retirarse del canal de inversión a la misma velocidad a la que cambia la tensión externa. Por tanto, aunque tengamos canal de inversión, éste no se modifica, por lo que no se apreciará el crecimiento de la capacidad en dicha región. Se observará lo que se muestra en la Figura 4.5.3 en línea discontinua:

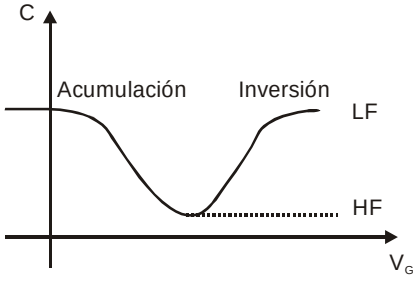


Figura 4.5.3 Capacidad vista entre los terminales de la estructura MIS en alta y baja frecuencia. La separación entre ambas regiones la define el tiempo que tardan los electrones del semiconductor en crear la lámina de inversión bajo el aislante.

RESUMEN

En este capítulo se ha descrito la parte fundamental del MOSFET: la estructura Metal-Aislante-Semiconductor. Se han estudiado las diferentes regiones en las que se puede encontrar este dispositivo de dos terminales y como evoluciona la carga del semiconductor al pasar de una región a otra. Se han definido las magnitudes eléctricas más significativas de la estructura. Se ha definido el valor de la tensión de puerta necesaria para que aparezca el canal de inversión de carga. Se ha encontrado una relación entre la carga almacenada en el semiconductor y la tensión externa aplicada a los terminales del dispositivo. Por último se ha encontrado una expresión para la capacidad de la estructura.

CUESTIONES Y PROBLEMAS

1. Considerar una estructura MIS donde el semiconductor es de silicio tipo P con una concentración de impurezas aceptadoras $N_A=10^{16}$ cm^{-3} y el aislante es SiO_2. ¿Cuál es el potencial V(x) en el semiconductor de una estructura MOS en inversión, calculado en dirección perpendicular a la superficie Si-SiO_2? ¿Cuál sería la concentración de electrones en el canal de inversión en esa misma dirección? ¿Cuál es la carga asociada al canal de inversión? ¿Y la carga almacenada en la región de vaciamiento?

   La solución a este problema depende del modelo que se utilice. Podemos tener en cuenta o no la cuantización en la lámina de inversión de carga. Hay que pensar que los electrones se encuentran confinados en una capa muy delgada, dentro de un pozo de potencial. Los estados permitidos de los electrones en ese pozo realmente están cuantizados. Aun despreciando la cuantización lo que sí deberíamos usar es el hecho de la degeneración del semiconductor en esa lámina (el nivel de Fermi sobrepasa con creces el fondo de la banda de conducción).

   A pesar de ello en este problema se propone estimar la concentración de electrones en el canal admitiendo que no existe ni cuantización ni degeneración.

   Para obtener la solución a este problema se propone resolver las siguientes cuestiones:

   a) Plantear la ecuación de Poisson en el semiconductor, desde la superficie con el óxido hasta el borde de la zona de carga espacial con la zona neutra, admitiendo que sólo existen impurezas ionizadas, despreciando los electrones y los huecos.

   b) Obtener el valor del campo eléctrico en la misma región.

c) Integrar una vez más para obtener la expresión del potencial.

d) A partir de la distribución de potencial calcular la concentración de electrones en esa región, $n(x)$. Integrar la expresión resultante a toda la región para extraer la carga de inversión por unidad de superficie Q_I.

e) Evaluar la anchura de la zona de carga espacial W y calcular la carga por unidad de superficie, Q_B, asociada a esta región.

f) Calcular la carga del semiconductor por unidad de superficie, Q_S, de acuerdo con el modelo utilizado en este capítulo.

g) Comprobar si esta Q_S, calculada con el modelo anterior, es igual a la suma $Q_I + Q_B$, obtenidas con el modelo propuesto en este problema. Identificar las posibles discrepancias.

h) Representar la densidad de electrones y la densidad de impurezas ionizadas a lo largo del semiconductor.

Considerar tres casos ($\Psi_S = 1.5\phi_F$, $\Psi_S = 2.0\phi_F$, $\Psi_S = 2.2\phi_F$) para analizar cómo evolucionan estas magnitudes con el paso de la zona de deplexión a inversión. Simular este transistor con PISCES y comparar el resultado con el obtenido analíticamente.

2. Considérese la estructura MIS de la **Figura P.1**. en el que el semiconductor de silicio está dopado con impurezas aceptadoras en concentración 10^{15} cm^{-3}. Se quiere comparar cuál es la densidad de electrones en fuerte inversión ($\psi_S = 2\phi_F$) y en débil inversión ($\psi_S = \phi_F$).Evaluar la densidad de electrones en estos dos casos justo en la superficie Si-SiO_2.

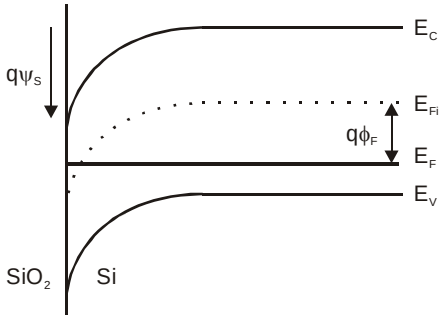


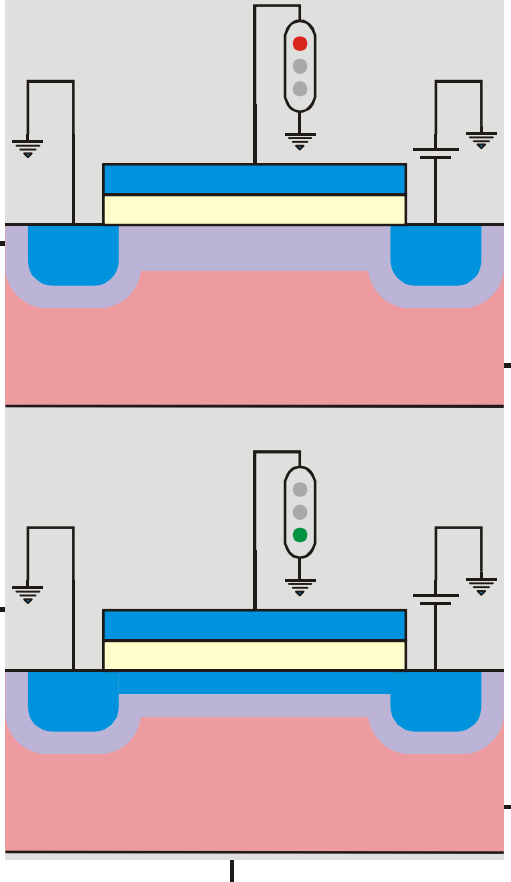
Figura P.1.

REFERENCIAS

[1] Y. P. Tsividis, *Operation and modeling of the MOS Transistor*. McGraw Hill, New York, 1987.

Capítulo 5

El transistor de Efecto Campo MOS (MOSFET).



Transistor MOSFET. Estados On y Off.

Índice

Objetivos

- Descripción del transistor completo, partiendo de la estructura MIS, con el fin de encontrar la característica corriente tensión.
- Encontrar las relaciones entre las corrientes que circulan por los tres terminales con las diferencias de potencial existentes entre ellos. Para ello se hará uso de un modelo de canal gradual y se considerarán campos eléctricos bajos.
- Encontrar un modelo lineal a partir de las relaciones no lineales obtenidas entre estas magnitudes. El objeto es simplificar el análisis de los circuitos que contengan a este tipo de dispositivos.
- Descripción de los modelos de gran señal y pequeña señal. Encontrar esos modelos lleva consigo sustituir el dispositivo por un circuito equivalente donde aparezcan exclusivamente elementos con característica I-V lineal.
- Mostrar como se utilizan estos modelos en el análisis de circuitos con MOSFETs. Análisis de un caso concreto: estudio de la respuesta en frecuencia del dispositivo.
- Introducción de otros mecanismos en los modelos. Efecto de campos eléctricos elevados en el canal, corriente subumbral y la corriente a través del sustrato.

Palabras Clave

MOSFET de enriquecimiento (normally off).
MOSFET de deplexión (normally on).
Canal de inversión de carga.
Tensión umbral.
Efecto “body”.
Característica I-V.
Región triodo.
Región de saturación.
Modulación de la longitud del canal.
Modelo de gran señal.
Modelo de pequeña señal.
Frecuencia para ganancia en corriente en cortocircuito unidad.
Efectos de canal corto.
Velocidad de saturación.
Conducción subumbral.
Corriente en el sustrato.

5.1 Estructura y símbolos de circuito

Una vez que se han asentado las bases de parte de este transistor, como es la estructura metal-aislante-semiconductor, ya se está en condiciones de abordar el estudio de la estructura completa. El dispositivo se muestra en la Figura 5.1.1.

MOSFET:
Transistor de efecto campo basado en la estructura MIS. La tensión aplicada a la puerta controla el espesor del canal, que une fuente y drenador, y en consecuencia controla la corriente que por él circula.

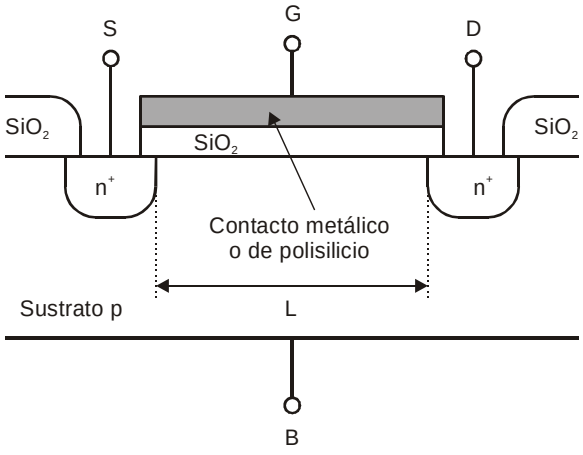


Figura 5.1.1 Estructura de un MOSFET de canal N. Definición de los contactos de fuente, S, puerta,G, drenador, D, y sustrato, B, y de la longitud del canal, L.

En esta estructura se distinguen cuatro terminales: puerta, G, y sustrato, B, definidos en la estructura MIS, y dos terminales más: drenador, D, y fuente, S. Se llaman así estos últimos porque entre drenador y fuente circula una corriente de forma que los portadores parten de la fuente y llegan al drenador. En esta figura el semiconductor del sustrato es silicio tipo P y los contactos de drenador y fuente son contactos óhmicos realizados sobre semiconductores N^+. El aislante es el propio óxido del semiconductor, SiO_2. Para que exista corriente entre fuente y drenador es necesario establecer un camino por el cual fluya la misma. Eso sólo es posible si se crea una lámina de inversión de carga de electrones que una las dos regiones N^+. Para ello hay que aplicar una tensión apropiada a la puerta.

Se pueden encontrar otras variedades de transistor MOSFET, distintas a las presentadas en la Figura 5.1.1, dependiendo del tipo de sustrato y si introducimos un canal entre fuente y drenador durante el proceso de fabricación del dispositivo. Se tienen los siguientes tipos de MOSFET:

MOSFET de canal N de enriquecimiento (normally off): el sustrato semiconductor es tipo P y no existe canal a menos que apliquemos la tensión necesaria a la puerta para que esto ocurra.

MOSFET de canal P de enriquecimiento (normally off): el sustrato semiconductor es tipo N y no existe canal a menos que apliquemos la tensión necesaria a la puerta para que esto ocurra.

MOSFET de canal N de deplexión (normally on): el sustrato semiconductor es tipo P y existe canal a tensión de puerta nula. Se puede hacer desaparecer al canal con la aplicación una tensión apropiada a la puerta.

MOSFET de canal P de deplexión (normally on): el sustrato semiconductor es tipo N y existe canal a tensión de puerta nula. Se puede hacer desaparecer al canal con la aplicación una tensión apropiada a la puerta.

Cada uno de estos cuatro transistores se puede representar con una pareja de símbolos como se muestran en la Figura 5.1.2. En el símbolo de la izquierda de cada pareja aparece una flecha que nos indica el sentido real de la corriente. En el símbolo de la derecha la punta de la flecha nos indica que el sustrato o el canal es tipo N (dependiendo hacia donde apunte la flecha).

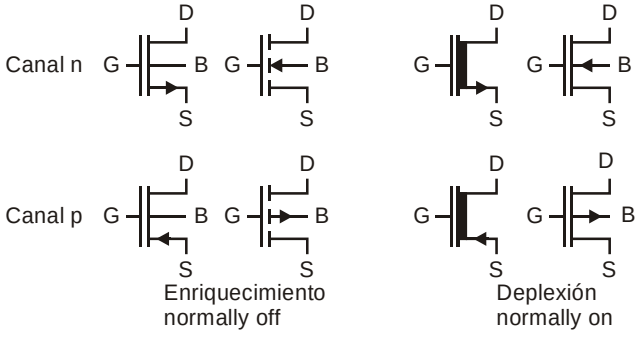


Figura 5.1.2 Símbolos del MOSFET empleados en circuitos. Hay cuatro modalidades dependiento del tipo del canal y de si existe canal o no bajo tensión nula.

5.2 Cálculo cualitativo de la característica corriente-tensión

Modo de operación

La idea de estos transistores, como cualquier otro de efecto campo, es poder controlar la corriente que circula por un canal semiconductor mediante la aplicación de un campo eléctrico. En el caso del MOSFET, el canal semiconductor lo constituye la lámina de inversión de carga próxima a la superficie con el óxido; la corriente fluye entre drenador y fuente y el campo eléctrico que controla el canal se aplica a la puerta. En el caso de un transistor canal N de enriquecimiento, como el de la Figura 5.2.1, para que exista canal es necesario aplicar una tensión positiva, V_{GS}. Por un lado repele los huecos de la superficie con el óxido

(creándose una zona de carga espacial) y por otro atrae electrones (formándose el canal propiamente dicho).

Tensión umbral:
Tensión aplicada a la puerta necesaria para que aparezca la lámina de inversión de carga y, por tanto, posibilite la conducción entre drenador y fuente.

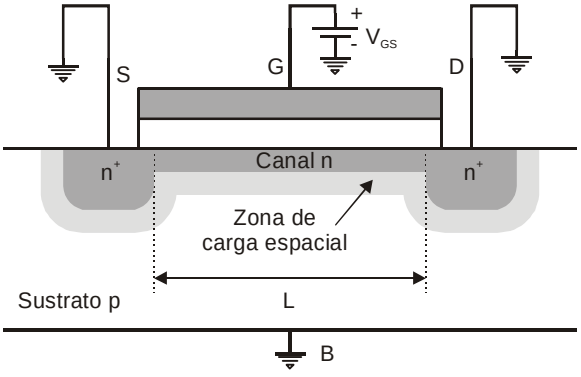


Figura 5.2.1 MOSFET con canal de inversión de carga bajo la puerta.

A la tensión aplicada a la puerta necesaria para que empiece a crearse el canal se le conoce con el nombre de tensión umbral, $V_{GS} = V_t$. Este es uno de los parámetros más importantes de este dispositivo pues es quien nos delimita los regímenes de conducción y no conducción en el MOSFET. Su estudio es tema prioritario, aunque antes de abordarlo es conveniente estudiar el efecto del terminal conectado al sustrato.

Efecto "body"

En el capítulo 4 se calculó la carga que existía en el semiconductor de una estructura MIS. En particular se calculó la carga de la zona de carga espacial cuando la estructura operaba en deplexión. Evaluando el valor de esa carga para el inicio de la región de inversión $(\psi_S = 2\phi_f)$ se tiene que la carga de la región de vaciamiento, Q_{bo}, vale:

$$Q_{bo} = -\sqrt{2qN_A\varepsilon_s 2\phi_f}\,. \qquad (5.1)$$

En la Figura 5.2.1 se observa que el canal está a cero voltios, pues están conectados a tierra el drenador y la fuente. El sustrato también se encuentra conectado a tierra. Si se observa exclusivamente el semiconductor se puede ver una unión pn con los dos extremos a tierra. Si ahora se aplicase una tensión negativa al sustrato (pues no interesa que circule corriente en esa unión pn) esta tensión se emplearía en aumentar la zona de carga espacial que existe debajo del canal. El nuevo valor de la carga, Q_b, dependería de la diferencia de potencial aplicada entre el sustrato y fuente, V_{SB}:

$$Q_b = -\sqrt{2q\,N_A\varepsilon_s(2\phi_f + V_{SB})}\,. \qquad (5.2)$$

Por regla general, la aplicación de tensiones al sustrato no se hace de manera intencionada. Lo más conveniente sería no añadir dificultad al estudio del dispositivo y conectar el sustrato al valor más

bajo de todas las tensiones que existan en el resto del mismo (con el fin de que la unión pn que se forma entre canal y sustrato no quede nunca polarizada en directo y no fluyan corrientes parásitas hacia el sustrato). En este caso, si se usan dos fuentes de alimentación para que funcione correctamente el dispositivo, una entre puerta y fuente, V_{GS}, para controlar la existencia de canal, y otra entre drenador y fuente, V_{DS}, para hacer circular una corriente entre el dispositivo, el sustrato debería conectarse a tierra, al igual que la fuente (Figura 5.2.2).

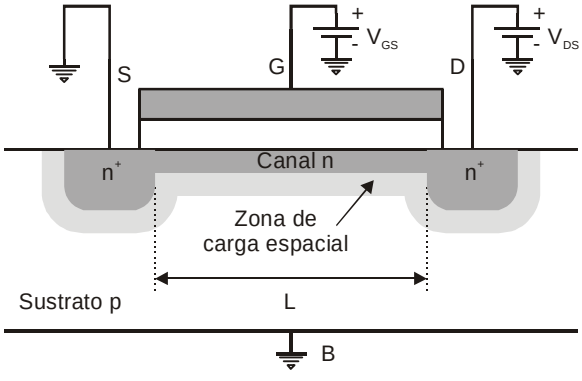


Figura 5.2.2 MOSFET de canal N con fuentes de alimentación en la puerta y el drenador.

Sin embargo, los dispositivos, al formar parte de un circuito, no se encuentran aislados sino que se encuentran conectados a otros elementos iguales o distintos. Consideremos a modo de ejemplo el circuito de la Figura 5.2.3, formado por dos transistores MOSFET de canal N.

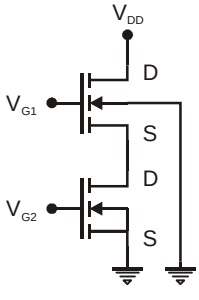


Figura 5.2.3 Combinación de dos MOSFETs en un circuito en el que se hace presente el efecto "body".

Las tensiones V_{DD}, V_{G1} y V_{G2} son positivas. Obsérvese como se ha conectado el sustrato de los dos transistores al valor más bajo de tensión que existe en el circuito, en este caso a tierra. Sin embargo, también se observa que mientras en el transistor "2" $V_{SB2} = 0$, en el transistor "1" $V_{SB1} > 0$. Ejemplos como este se pueden encontrar en muchos circuitos con MOSFETs. Para hacer más general el estudio de

este transistor se deberá tener en cuenta que la diferencia de potencial entre fuente y sustrato es mayor o igual que cero ($V_{SB} \geq 0$).

Cálculo de la tensión umbral V_t

En el capítulo 4 se obtuvo una relación entre la tensión aplicada a la puerta, V_G, con el potencial de superficie, ψ_S , y la carga almacenada en el semiconductor de un MIS, Q_S . Si particularizamos ahora para el inicio de la región de inversión, $V_{GS} \equiv V_G = V_t$, la carga en el semiconductor será debida exclusivamente a la carga de la zona de vaciamiento, $Q_S = Q_b$, y el potencial de superficie tomará el valor $\psi_S = 2\phi_F$. La relación queda por tanto:

$$\begin{aligned} V_t &= \frac{\phi_{ms}}{q} + 2\phi_f - \frac{Q_b}{C_{ox}} - \frac{Q_{ss}+Q_{ox}}{C_{ox}} = \\ &= \frac{\phi_{ms}}{q} + 2\phi_f - \frac{Q_{bo}}{C_{ox}} - \frac{Q_{ss}+Q_{ox}}{C_{ox}} - \frac{Q_b - Q_{bo}}{C_{ox}}, \end{aligned} \tag{5.3}$$

$$V_t = V_{t0} + \gamma\left(\sqrt{2\phi_f + V_{SB}} - \sqrt{2\phi_f}\right),$$

$$\gamma = \frac{1}{C_{ox}}\sqrt{2q\varepsilon_s N_A}, \quad C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}},$$

donde se ha definido V_{t0} como la tensión umbral para V_{SB}=0. Valores típicos para el parámetro γ y la capacidad del óxido por unidad de superficie C_{OX} son: $\gamma = 0.5$ V$^{1/2}$, $C_{ox} = 3.5\times10^{-4}$ pF μm^{-2}.

La tensión umbral V_{t0} se puede controlar añadiendo impurezas aceptadoras al semiconductor. Este es uno de los parámetros de diseño más relevantes del transistor por lo que su control resulta primordial. Para espesores del óxido de unas 0.1 μm V_{t0} puede oscilar entre 0.5 y 1.5 V.

Si las impurezas se introducen en una capa muy delgada próxima a la superficie del semiconductor la variación de la tensión umbral vendrá dada por:

$$\Delta V_t = \frac{Q_i}{C_{ox}}, \tag{5.4}$$

donde Qi es la carga introducida por unidad de superficie. Si las nuevas impurezas penetran una distancia considerable en el semiconductor habría que recalcular la expresión (5.3) con los nuevos dopados.

También se puede controlar la tensión umbral añadiendo impurezas donadoras, es decir, implantado un canal tipo N, con lo que tendríamos un MOSFET canal N de deplexión. En este caso se pueden conseguir valores de tensión umbral, V_{t0}, entre –1 y –4 V.

5.3 Característica de gran señal

Modelo cualitativo

Consideremos la estructura MOSFET polarizada con las fuentes de alimentación que se muestran en la Figura 5.3.1. Para obtener las relaciones corriente tensión en un MOSFET habrá que relacionar la corriente que circula entre drenador y fuente con las diferencias de potencial V_{GS}, V_{DS} y V_{SB}. No existen otras corrientes en el dispositivo, o al menos apreciables, pues en el terminal de puerta nos encontramos un aislante y en el terminal de sustrato una unión polarizada en inverso.

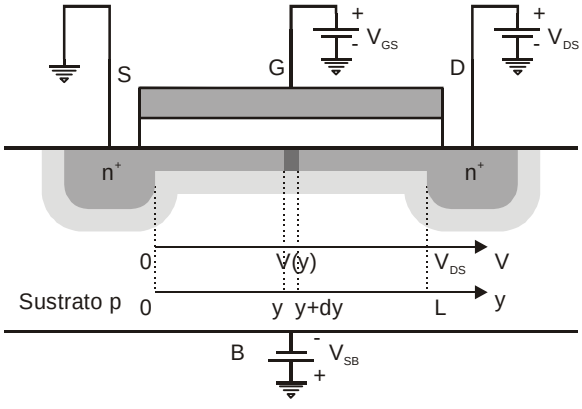


Figura 5.3.1 MOSFET de canal N. Definición del potencial $V(y)$ en un elemento de longitud del canal situado a una distancia y de la fuente.

En primer lugar se va a fijar la tensión V_{GS} de forma que exista canal de conducción, y se va a estudiar qué ocurre cuando se varía la tensión V_{DS}. La existencia de una diferencia de potencial entre drenador y fuente hace que la diferencia de potencial entre el metal de puerta y cualquier punto del canal sea diferente según la posición en el canal. Así, en el extremo de fuente la diferencia de potencial puerta-canal coincidirá con V_{GS}, pero en el extremo de drenador esa diferencia de potencial será igual a $V_{GD} = V_{GS} - V_{DS}$. Si llamamos $V(y)$ a la tensión de un punto y del canal, con $V(y)$ variando entre $V(0) = 0$ y $V(L) = V_{DS}$, la diferencia de potencial puerta-canal en ese punto y será $V_{GS} - V(y)$. Esto significa que si la tensión puerta canal varía con la posición también variará el tamaño del canal de conducción. Nos podemos encontrar con las situaciones que se muestran en la Figura 5.3.2 dependiendo del valor de la tensión V_{DS}:

(a) Para valores bajos de la tensión V_{DS} el canal se puede considerar uniforme.

(b) Para valores superiores sí es apreciable la diferencia de espesor del canal entre fuente y drenador.

(c) Si se sigue aumentando la tensión V_{DS} llegará un momento que se agote el canal. Eso ocurre cuando se dé la condición $V_{GD} = V_t$. Al valor de la tensión V_{DS} que permite que esto ocurra se le denomina V_{DSsat}.

(d) Para valores más altos de V_{DSsat}. El agotamiento del canal tendrá lugar en lugares cada vez más alejados del drenador.

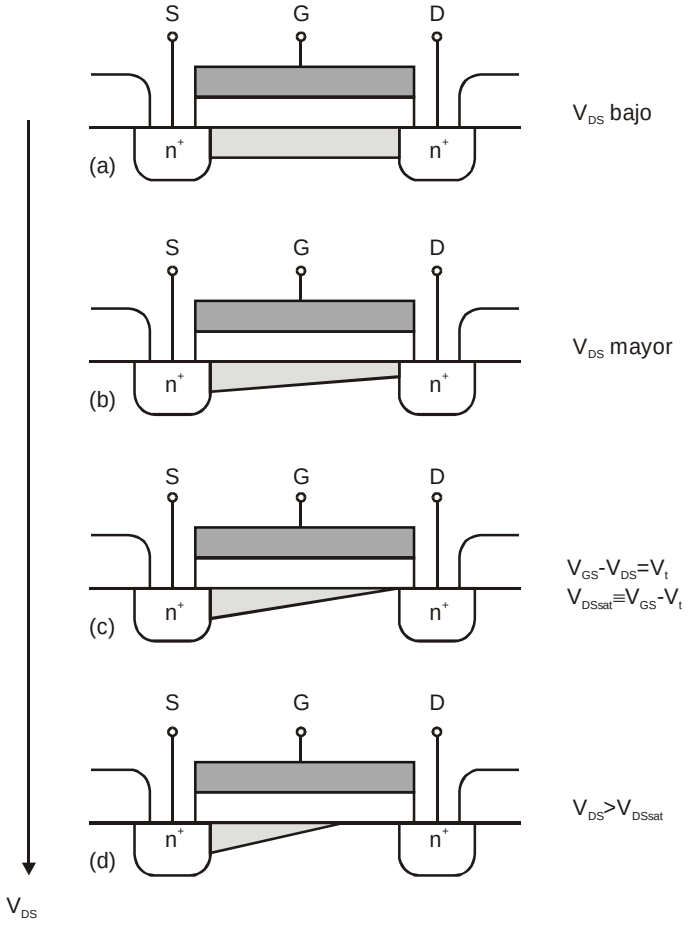


Figura 5.3.2 Evolución del canal de un MOSFET cuando se modifica la tensión V_{DS} para un valor fijo de la tensión V_{GS}.

La corriente que circula por el canal también depende de cada una de las situaciones anteriores (Figura 5.3.3).

(a) En este caso el canal se comporta como una resistencia cuya forma geométrica corresponde a un paralelepípedo. La relación corriente tensión será lineal.

(b) En esta región tenemos de nuevo un trozo de semiconductor, por lo que tendremos un comportamiento resistivo. Sin embargo, la sección

de este semiconductor en uno de sus extremos se va haciendo más pequeña de forma progresiva. Eso significa que la resistencia que presenta es mayor a medida que aumenta la tensión V_{DS}. En consecuencia, la pendiente de la curva I-V irá disminuyendo hasta anularse.

(c) Punto de agotamiento del canal en el drenador. Es el punto donde se anula la pendiente de la curva I-V.

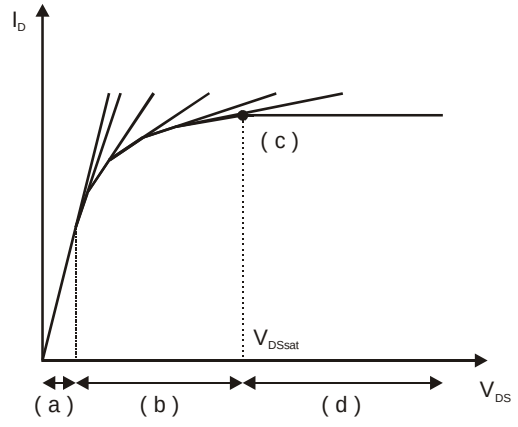


Figura 5.3.3 Evolución de la corriente que circula por el canal de un MOSFET para los casos definidos en la Figura 5.3.2.

(d) A partir de ese punto el agotamiento del canal se desplaza hacia la fuente. Se aprecia una región con carga de inversión con forma triangular en la que existe una diferencia de potencial entre sus extremos igual a V_{DSsat} (condición de agotamiento del canal). Si admitiésemos que la longitud de esa cuña no disminuye mucho con la tensión V_{DS}, al no variar la forma de manera apreciable y al tener siempre aplicada la misma diferencia de potencial, tendríamos la misma corriente. Esto realmente no es cierto pues la longitud de esa región de carga de inversión se hace más pequeña con lo que su resistencia disminuiría y la pendiente de la curva I-V se haría mayor. Esto se conoce como efecto de la longitud efectiva del canal y lo trataremos más adelante. Podrían hacerse dos objeciones al cálculo de la corriente en esta región: (i) Que la corriente se anule por agotarse los portadores en parte del canal. (ii) Que la velocidad de los electrones se haga infinita para poder mantener una corriente finita con una densidad de electrones nula. A la primera objeción se responde diciendo que si en parte del canal hay portadores disponibles y existe una diferencia de potencial entre sus extremos, necesariamente habrá un flujo de electrones. Este flujo de portadores no podrá detenerse de forma brusca en el extremo donde se agota el canal a menos que exista otra fuerza opuesta. Realmente hay electrones atravesando todo el canal, por lo que el canal no se elimina

por completo. Tendrá el tamaño necesario para permitir ese flujo de electrones. Esto responde también a la segunda objeción, pues si existe canal ya no es necesario hacer tender la corriente a infinito, hecho que no sería justificable físicamente.

Modelo cuantitativo

Una vez descrito el comportamiento cualitativo del MOSFET vamos a encontrar expresiones analíticas para la característica corriente tensión. Para ello vamos a dividir el canal en unidades de longitud infinitesimales de manera que podamos considerar uniforme el espesor del canal en cada elemento (Figura 5.3.1). Sea *dy* la longitud de ese canal infinitesimal, e *y* su distancia a la fuente. Como hipótesis se va a considerar que la carga inducida en el canal por unidad de superficie se puede expresar como:

$$Q_I(y) = C_{ox}(V_{GS} - V(y) - V_t)\,. \tag{5.5}$$

El significado de esta expresión es que toda la tensión aplicada entre la puerta y el canal, $(V_{GS} - V(y))$, que supere la tensión umbral V_t se emplea en aumentar la carga de la lámina de inversión. Estrictamente esto no es cierto, porque parte de la tensión aplicada se emplea también en aumentar la carga de la zona de vaciamiento. Sin embargo, para establecer un primer contacto con este dispositivo se asumirá esta hipótesis.

La resistencia de este elemento del canal vendrá dada por:

$$dR = \frac{dy}{W\,\mu_n Q_I(y)}, \tag{5.6}$$

donde W es la profundidad del dispositivo y μ_n es la movilidad de los electrones.

La diferencia de potencial en los extremos de ese elemento será

$$dV = I_D\,dR = \frac{I_D}{W\,\mu_n Q_I(y)}dy\,, \tag{5.7}$$

donde I_D es la corriente que atraviesa el canal. Integrando la expresión anterior a lo largo del canal se obtiene

$$\int_0^L I_D\,dy = \int_o^{V_{DS}} W\,\mu_n Q_I(y)\,dV = \int_0^{V_{DS}} W\,\mu_n C_{ox}(V_{GS} - V - V_t)\,dV,$$

$$I_D = \frac{k'}{2}\frac{W}{L}\left[2(V_{GS} - V_t)V_{DS} - V_{DS}^2\right], \tag{5.8}$$

$$k' = \mu_n C_{ox} = \mu_n \frac{\varepsilon_{ox}}{t_{ox}}.$$

Esta función es una parábola que presenta un máximo para $V_{DS} = V_{GS} - V_t$, justo el valor del agotamiento del canal en el drenador. A partir de ese valor no tiene sentido usar la expresión parabólica pues eso implicaría una disminución de la corriente. Hemos visto que a partir del

valor de la tensión de agotamiento la corriente se mantiene constante, tomando el valor del máximo. La expresión de la corriente en todo el rango de tensiones quedaría de la forma:

$$I_D = \frac{k'}{2}\frac{W}{L}\left[2(V_{GS} - V_t)V_{DS} - V_{DS}^2\right] \qquad V_{DS} < V_{GS} - V_t,$$
$$I_D = \frac{k'}{2}\frac{W}{L}(V_{GS} - V_t)^2 \qquad V_{DS} \geq V_{GS} - V_t. \tag{5.9}$$

Cuando el transistor se encuentra en la primera región (tensiones inferiores al valor de agotamiento) se dice que trabaja en la región triodo y cuando se encuentra en la segunda se dice que trabaja en saturación (por el valor constante de la corriente).

Región triodo:
El canal no se ha agotado en ningún punto del trayecto entre drenador y fuente.

Región lineal:
Caso particular de la región triodo donde el espesor del canal se puede considerar uniforme a lo largo del mismo (en algunas ocasiones se utilizan como sinónimas esta región y la triodo).

Región de saturación:
El canal de inversión desaparece en las proximidades del contacto de drenador.

Corrección por la modulación del canal

Se ha mencionado anteriormente que la corriente no es exactamente constante una vez que aparece el agotamiento del canal en la región de drenador. Para tener en cuenta la longitud real del canal se debería sustituir el parámetro longitud física, L, por longitud efectiva, $L_{eff} = L - x_d$, donde x_d es la distancia que separa el punto de agotamiento en el canal del drenador (Figura 5.3.4). Introduciendo esta variable, la corriente en la región de saturación quedaría:

$$I_D = \frac{k'}{2}\frac{W}{L_{eff}}(V_{GS} - V_t)^2 \qquad V_{DS} \geq V_{GS} - V_t. \tag{5.10}$$

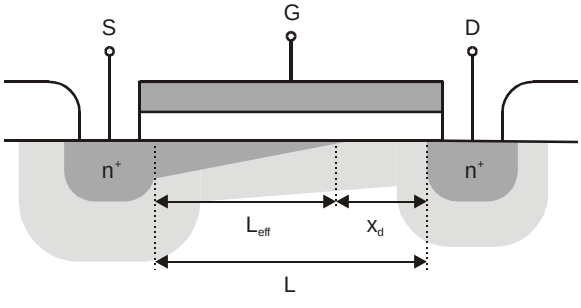


Figura 5.3.4 Definición de la longitud efectiva del canal en un MOSFET trabajando en la región de saturación.

Esta expresión presenta un problema: se desconoce cuál es la variación de L_{eff} con la tensión V_{DS}. Sería muy útil encontrar un parámetro que modelara este efecto sin tener que encontrar la forma de la función $L_{eff} = L_{eff}(V_{DS})$. Para ello se calcula la pendiente de la corriente en la región de saturación:

$$\frac{\partial I_D}{\partial V_{DS}} = -\frac{k'}{2}\frac{W}{L_{eff}}(V_{GS} - V_t)^2\frac{dL_{eff}}{dV_{DS}} = \frac{I_d}{L_{eff}}\frac{dx_d}{dV_{DS}}. \tag{5.11}$$

Si en la expresión anterior agrupamos en un término todo lo que depende de la corriente y en el otro término lo que no depende de ella encontramos:

$$\frac{I_D}{\dfrac{\partial I_D}{\partial V_{DS}}} = L_{eff}\left(\frac{dx_d}{dV_{DS}}\right)^{-1} \quad . \tag{5.12}$$

Definiendo los parámetros tensión Early, V_A, y λ como:

$$V_A \equiv L_{eff}\left(\frac{dx_d}{dV_{DS}}\right)^{-1} \qquad \lambda \equiv \frac{1}{V_A} \approx (0.05\text{-}0.005\ \mathrm{V}^{-1}), \tag{5.13}$$

podemos escribir la corriente en la región de saturación empleando este último parámetro:

$$I_D = \frac{k'}{2}\frac{W}{L}(V_{GS} - V_t)^2(1 + \lambda V_{DS}) \, . \tag{5.14}$$

De acuerdo con la definición de la tensión Early, V_A es un parámetro que es independiente de la corriente I_D. Se puede comprobar que es un parámetro único para todo el dispositivo sin más que evaluar el cociente $I_D/(\partial I_D/\partial V_{DS})$ para cualquier curva $I_D - V_{DS}$, tal y como se observa en la Figura 5.3.5. El resultado de ese cociente es realmente la base de todos los triángulos mostrados en la figura. Como se observa esa base es común a todos ellos. Los triángulos se construyen a partir de la extrapolación de las curvas I-V en saturación hacia valores negativos de la tensión.

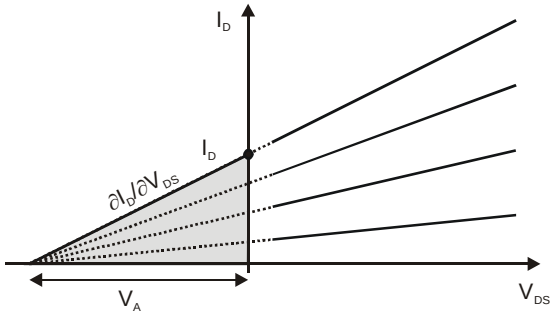


Figura 5.3.5 Significado de la tensión Early. Base común a todos los triángulos construidos extrapolando las curvas I-V en saturación.

Ejemplo 5.1

El modelo empleado para calcular la característica I-V en un MOSFET considera que una variación del potencial $V(y)$ en un punto y del canal sólo afecta a la carga del canal de inversión y no a la carga de la región de vaciamiento, que coexiste debajo del canal,

según se desprende de la ecuación (5.5). En este ejemplo se pretende deducir una nueva expresión para la corriente $I_D = I_D(V_{DS}, V_{GS})$ en un MOSFET incluyendo los efectos de la zona de carga espacial.

Solución.

Una forma más correcta de calcular dicha corriente es considerar que la carga de deplexión también se ve afectada por ese potencial $V(y)$:

$$V_t = \frac{\phi_{ms}}{q} - \frac{Q_{ss}+Q_{ox}}{C_{ox}} + 2\phi_f - \frac{Q_b}{C_{ox}} = V_{FB} + 2\phi_F + \gamma\sqrt{2\phi_F + V(y)},$$
$$V_{FB} = \frac{\phi_{ms}}{q} - \frac{Q_{ss}+Q_{ox}}{C_{ox}}. \qquad (5.15)$$

De esta forma la carga en inversión se puede escribir como:

$$Q_I(y) = -C_{ox}(V_{GS} - V_{FB} - 2\phi_F - V(y) - \gamma\sqrt{2\phi_F + V(y)}). \qquad (5.16)$$

Para calcular la corriente se sustituye esta nueva expresión de $Q_I(y)$ en (5.6):

$$dR = \frac{dy}{\mu_n W C_{ox}[V_{GS} - V_{FB} - 2\phi_F - V(y) - \gamma\sqrt{2\phi_F + V(y)}]},$$

$$\int_0^L I_D dy = \int_0^{V_{DS}} \mu_n C_{ox} W[V_{GS} - V_{FB} - 2\phi_F - V(y) - \gamma\sqrt{2\phi_F + V(y)}]dV, \qquad (5.17)$$

$$I_D = \mu_n C_{ox}\frac{W}{L}\left[[(V_{GS} - V_{FB} - 2\phi_F) - \frac{V_{DS}}{2}]V_{DS} - \frac{2}{3}\gamma[(2\phi_F - V_{DS})^{\frac{3}{2}} - (2\phi_F)^{\frac{3}{2}}]\right].$$

Si se desarrolla en serie el siguiente término:

$$(2\phi_F + V_{DS})^{\frac{3}{2}} = (2\phi_F)^{\frac{3}{2}} + \frac{3}{2}(2\phi_F)^{\frac{1}{2}}V_{DS} + \frac{3}{8}(2\phi_F)^{-\frac{1}{2}}V_{DS}^2 + \ldots \qquad (5.18)$$

y admitimos que trabajamos a tensiones tales que $V_{DS} < 2\phi_F$ podemos quedarnos con los dos primeros términos del desarrollo, con lo quedaría una corriente igual a la que se ha obtenido en (5.9). Para tensiones V_{DS} más elevadas deberemos incluir el tercer término del desarrollo en serie con lo que la corriente tomará la forma:

$$I_D = \frac{K'}{2}\frac{W}{L}\left[2(V_{GS} - V_t)V_{DS} - (1+\delta)V_{DS}^2\right] \quad V_{DS} < V_{DSsat},$$

$$I_D = \frac{K'}{2}\frac{W}{L}\left[\frac{(V_{GS} - V_t)^2}{(1+\delta)}\right] \quad V_{DS} > V_{DSsat}, \qquad (5.19)$$

$$\delta = \frac{\gamma}{2\sqrt{2\phi_F}} \quad V_{DSsat} = \frac{(V_{GS} - V_t)}{(1+\delta)}.$$

5.4 Curvas características del MOSFET

En un MOSFET se puede trabajar con dos tipos de curvas I-V (Figura 5.4.1). En una se representa la corriente de drenador, I_D, en función de la tensión drenador-fuente, V_{DS}, manteniendo como parámetro la tensión puerta fuente V_{GS}. Se puede ver como al aumentar V_{GS} la corriente también lo hace. En esa gráfica se pueden distinguir las regiones óhmica o triodo, la de saturación y la región de corte, caracterizada esta última porque no circula corriente por el transistor. En la otra gráfica se hace una representación de la corriente de drenador en función de la tensión puerta fuente, V_{GS}, en la región de saturación. Dependiendo del valor de la tensión fuente sustrato se tendrán diferentes valores para la tensión umbral, y por tanto una menor corriente si aumenta V_{SB}.

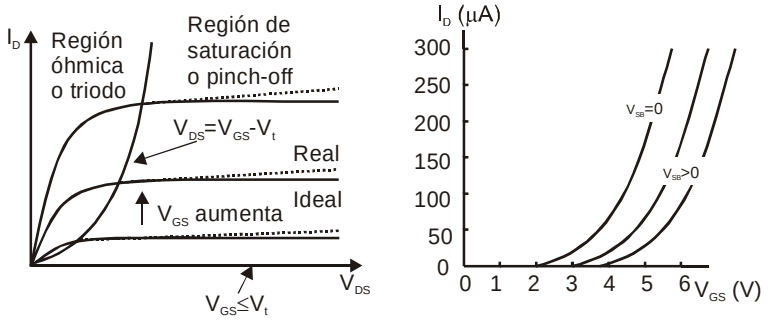


Figura 5.4.1 Curvas I_D-V_{DS} para V_{GS} constante y curvas I_D-V_{GS} en la región de saturación de un MOSFET canal N de enriquecimiento.

Para un transistor NMOS de deplexión las curvas características tienen el mismo aspecto que el de enriquecimiento salvo que hay un desplazamiento de la tensión umbral (Figura 5.4.2). Hay conducción para $V_{GS} = 0$ e incluso para valores negativos de esta tensión, hasta que se alcance la tensión umbral, también negativa. En el caso de transistores de canal P habría que cambiar el signo tanto a la corriente como a las tensiones del dispositivo.

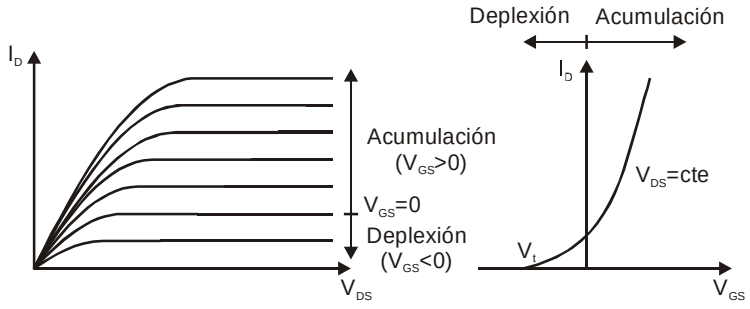


Figura 5.4.2 Diferentes curvas características de un MOSFET canal N de deplexión: Existe conducción para $V_{GS} \geq V_t$, donde ahora $V_t < 0$.

5.5 Modelos de circuito del MOSFET

Modelo de gran señal

Cuando el MOSFET forma parte de un circuito es conveniente trabajar con un modelo que simplifique el análisis de ese circuito. En primer lugar debemos conocer, o asumir y comprobar posteriormente, en qué región trabaja el dispositivo. En el caso de trabajar en la región de saturación el modelo sería el de la Figura 5.5.1.

El modelo refleja el modo de operación del MOSFET. Presenta una resistencia infinita en el terminal de puerta como consecuencia del aislante. Dispone de una fuente de corriente entre fuente y drenador dependiente de la diferencia de potencial entre puerta y fuente. Si se quiere incluir el efecto Early habría que hacer uso de la expresión (5.14), que incluye al parámetro λ. Si se trabaja en la región triodo habría que sustituir la expresión de la corriente de la fuente dependiente que aparece en el modelo por la correspondiente a esta región triodo (5.9).

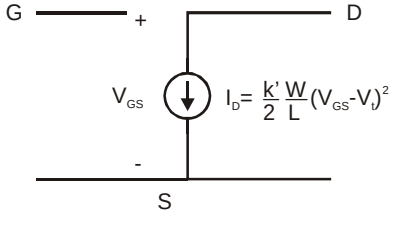


Figura 5.5.1 Modelo de gran señal de un MOSFET.

Modelo de pequeña señal en la región de saturación

Un transistor en general tiene interés cuando se utiliza en aplicaciones dinámicas, es decir, cuando sus variables cambian con el tiempo. Resulta primordial conocer, además de la relación entre las variables estáticas, la relación entre las variaciones de estas mismas variables.

Cuando se emplea el transistor MOSFET como amplificador interesa elegir la zona de saturación y no la región triodo. En primer lugar porque en la zona de saturación la corriente de drenador depende de una sola variable, la tensión de control V_{GS}, mientras que en la región triodo depende además de la tensión V_{DS}. Por otro lado, al ser la característica I–V no lineal, la distorsión que se obtenga a la salida del amplificador será mayor en la zona triodo que en la de saturación. Por estas razones nos vamos a centrar en la región de saturación.

Analicemos la expresión de la corriente en saturación desde un punto puramente formal y veamos qué ocurre cuando las variables que en ella aparecen cambian con el tiempo.

La corriente

$$I_D = \frac{k'}{2}\frac{W}{L}(V_{GS} - V_t)^2(1 + \lambda V_{DS}) \tag{5.20}$$

es función de tres variables: V_{GS}, V_{DS} y V_{SB}. La última está implícita en la tensión umbral. Ya vimos que la tensión V_{SB} no tiene porqué ser cero, lo mismo que tampoco tiene que ser cero su variación temporal. Es como si el sustrato actuara de segunda puerta en el MOSFET. Aunque de forma intencionada solo modifiquemos externamente la tensión de control V_{GS}, por influencia del resto del circuito del que forma parte este transistor, las tensiones V_{DS} y V_{SB} también pueden variar. Consideremos por tanto a estas variables como la suma de su valor estático más un incremento:

$$\begin{aligned} V_{GS} &= V_{GS0} + \Delta V_{GS}, \\ V_{DS} &= V_{DS0} + \Delta V_{DS}, \\ V_{SB} &= V_{SB0} + \Delta V_{SB}. \end{aligned} \tag{5.21}$$

La ecuación de la corriente (5.20) se puede desarrollar en serie de Taylor y expresarse en función de los incrementos anteriores:

$$\Delta I_D = I_D - I_{D0} = \frac{\partial I_D}{\partial V_{GS}}\Delta V_{GS} + \frac{\partial I_D}{\partial V_{DS}}\Delta V_{DS} + \frac{\partial I_D}{\partial V_{SB}}\Delta V_{SB} + \ldots, \tag{5.22}$$

donde solo se han escrito los términos de primer orden. Si las variaciones de tensión que ahí aparece son pequeñas nos podríamos quedar exclusivamente con estos términos del desarrollo. Tendríamos una relación puramente lineal entre las variaciones de las distintas variables, donde las derivadas parciales serían parámetros. Identificando las variaciones de las variables con su notación en minúscula tendríamos:

$$i_d = g_m v_{gs} + g_D v_{ds} + g_{mb} v_{sb}, \tag{5.23}$$

donde las transconductancias g_m, g_{mb} y la conductancia g_D se definen como:

$$g_m \equiv \frac{\partial I_D}{\partial V_{GS}},$$
$$g_{mb} \equiv \frac{\partial I_D}{\partial V_{SB}}, \qquad (5.24)$$
$$g_D \equiv \frac{\partial I_D}{\partial V_{DS}}.$$

La ecuación (5.23) tiene un equivalente de circuito directo: la combinación de tres elementos en paralelo como se muestra en la Figura 5.5.2.

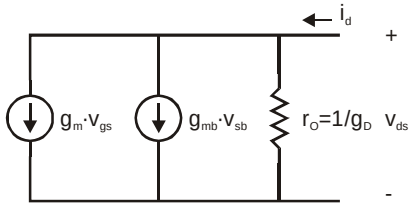


Figura 5.5.2 Circuito eléctrico equivalente que modela la ecuación (5.23).

Desde el punto de vista del análisis de circuitos lo que nos estamos planteando se muestra en la Figura 5.5.3. Realmente nos estamos preguntando por la corriente que circula por el drenador con una configuración de fuentes de corriente continua (que representan el comportamieno estático) y fuentes de corriente alterna (que representan las variaciones de tensión).

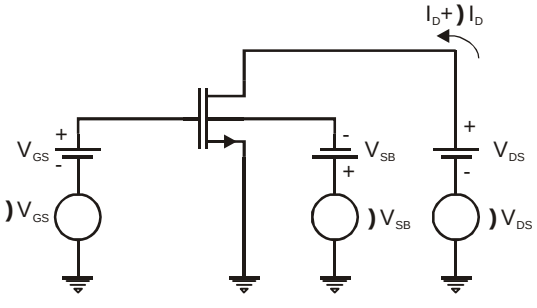


Figura 5.5.3 Combinación de fuentes de corriente continua y alterna en un circuito con un MOSFET.

El resultado que se ha encontrado en (5.23), donde aparece una relación lineal entre las distintas variaciones de las tensiones y corriente (siempre que esas variaciones sean pequeñas, lo que hace posible introducir el concepto de pequeña señal), nos permite analizar este circuito aplicando el principio de superposición. Eso es posible porque también se dispone de un modelo lineal para estudiar al transistor en condiciones de gran señal. Antes de hacer ésto se debe encontrar el

modelo de pequeña señal, aunque ya está representado parte del mismo en la Figura 5.5.2.

Dicho modelo debe relacionar variaciones de la corriente de salida con variaciones de señal a la entrada ΔV_{GS}, ΔV_{SB}. En él no deben intervenir las variables estáticas o dc aunque sus elementos dependan intrínsecamente de ellas. Los parámetros que deben incluirse son los siguientes:

Transconductancias: reflejan las relaciones entre las variaciones de las fuentes de tensión V_{GS} y V_{SB} con la corriente I_D.

Resistencia de salida: indica la relación entre las variaciones de tensión y corriente a la salida. Es consecuencia directa del efecto de la modulación de la longitud del canal o efecto Early.

Capacidades: afectan al comportamiento en frecuencia e introducen retardos. Incluyen las regiones de la estructura donde aparece algún almacenamiento de carga eléctrica.

Resistencias parásitas asociadas a los contactos: reflejan los elementos resistivos que se encuentra la corriente a su paso por las diferentes regiones del dispositivo.

El valor de cada uno de estos elementos se comenta a continuación. Posteriormente se irán combinando cada uno de estos elementos para obtener el circuito equivalente del transistor.

- Cálculo de la transconductancia g_m. Haciendo uso de las definiciones (5.24) se obtiene:

$$g_m = \frac{\partial I_D}{\partial V_{GS}} = k' \frac{W}{L} (V_{GS} - V_t)(1 + \lambda V_{DS}) . \tag{5.25}$$

  En el caso que $\lambda V_{DS} << 1$ se puede aproximar la transconductacia por

$$g_m \approx k' \frac{W}{L} (V_{GS} - V_T) = \sqrt{2k' \frac{W}{L} I_D} . \tag{5.26}$$

- Cálculo de la transconductancia g_{mb}:

$$g_{mb} = \frac{\partial I_D}{\partial V_{BS}} = -k' \frac{W}{L} (V_{GS} - V_t)(1 + \lambda V_{DS}) \frac{\partial V_t}{\partial V_{BS}} . \tag{5.27}$$

  Derivando la tensión umbral dada en (5.3) se obtiene:

$$\frac{\partial V_t}{\partial V_{BS}} = - \frac{\gamma}{2\sqrt{2\phi_f + V_{SB}}} \equiv -\chi . \tag{5.28}$$

  En el caso que $\lambda V_{DS} << 1$ se puede aproximar por:

$$g_{mb} = \frac{k'\frac{W}{L}(V_{GS} - V_t)(1 + \lambda V_{DS})\gamma}{2\sqrt{2\phi_f + V_{SB}}} \approx \frac{\gamma\sqrt{k'\frac{W}{L} I_D}}{\sqrt{2(2\phi_f + V_{SB})}} . \qquad (5.29)$$

Comparando las expresiones (5.26) y (5.29) se obtiene la siguiente relación entre las transconductancias:

$$\frac{g_{mb}}{g_m} = \frac{\gamma}{2\sqrt{2\phi_f + V_{SB}}} = \chi \ . \qquad (5.30)$$

El parámetro χ, definido en (5.28), puede tomar valores típicos entre 0.1 y 0.3. Este parámetro se obtiene también como resultado de dividir la capacidad asociada a la zona de carga espacial que existe debajo de la capa de inversión, C_{js}, (para calcularla se hace uso de la ecuación (4.16)) y la capacidad del óxido de puerta, C_{OX}:

$$\chi = \frac{C_{js}}{C_{ox}};\ C_{js} = \sqrt{\frac{q\,\varepsilon_S\,N_A}{2(2\phi_f + V_{SB})}} . \qquad (5.31)$$

- Cálculo de la resistencia de salida r_O:

$$r_o = \left(\frac{\partial I_D}{\partial V_{DS}}\right)^{-1} = \frac{Leff}{I_D}\left(\frac{dx_d}{dV_{DS}}\right)^{-1} = \frac{1}{\lambda I_D} = \frac{V_A}{I_D} . \qquad (5.32)$$

Como se observa, r_O es función de la tensión Early.

- Cálculo de las capacidades. En este dispositivo existen varias regiones que introducen efectos capacitivos:
  - ✓ Las zonas de carga espacial de las uniones drenador-sustrato y fuente sustrato. Las expresiones de estas capacidades, C_{sb} y C_{db} respectivamente, corresponden a las de dos uniones polarizadas en inverso:

$$C_{sb} = \frac{C_{sb0}}{\sqrt{1 + \frac{V_{SB}}{\psi_o}}}$$
$$C_{db} = \frac{C_{db0}}{\sqrt{1 + \frac{V_{DB}}{\psi_o}}} \qquad (5.33)$$

donde C_{sb0} y C_{db0} son las capacidades a tensión cero, V_{SB} y V_{DB} son las diferencias de potencial en las uniones fuente-sustrato y drenador-sustrato respectivamente y ψ_0 es el potencial barrera de esas uniones.

  - ✓ Las regiones de interconexiones de puerta fuera de la zona activa del transistor. Dan lugar a una capacidad parásita entre el material de puerta y el sustrato, C_{gb}.

Los valores típicos de esta capacidad oscilan entre 0.04 y 0.15 fF por micrómetro cuadrado de interconexión.

- ✓ Capacidad de la estructura MOS. Recordemos que en la estructura MIS la capacidad de la estructura tanto en acumulación como en inversión coincidía con la capacidad del óxido. En el caso del MOSFET existe canal trabajando tanto en la región triodo como en la de saturación. Sin embargo, en saturación la carga de inversión no se reparte por todo el canal, sino que se encuentra más cerca de la fuente. Hay que distinguir por consiguiente entre la capacidad asociada a la zona puerta-fuente, C_{gs}, y la capacidad puerta-drenador, C_{gd}. También hay que diferenciar si se trabaja en la región triodo o saturación.

  En la región óhmica el canal tiene un espesor prácticamente uniforme por lo que las dos capacidades serán iguales y su valor será el de la mitad de la capacidad del óxido: $C_{gs} = C_{gd} = (C_{OX}WL)/2$.

  En saturación la región de drenador no contribuye con efectos capacitivos pues no existe carga en esa zona. Únicamente habrá que tener en cuenta la capacidad asociada al pequeño solapamiento que existe entre el óxido de puerta y la región N^+ de drenador $(C_{gd} \approx 0)$.

  Para calcular la capacidad C_{gs} habrá que ver como se modifica la carga almacenada en el canal, Q_T, cuando se produce una variación de la tensión V_{GS}.

$$C_{gs} = \frac{\partial Q_T}{\partial V_{GS}}, \tag{5.34}$$

  donde, de acuerdo con (5.5), la carga almacenada en el canal se puede expresar como:

$$\begin{aligned} Q_T &= W C_{ox} \int_0^L (V_{GS} - V(y) - V_t)dy \\ &= \frac{W^2 C_{ox}^2 \mu_n}{I_D} \int_0^{V_{GS}-V_t} (V_{GS} - V - V_t)^2 dV = \\ &= \frac{2}{3} WL C_{ox} (V_{GS} - V_t). \end{aligned} \tag{5.35}$$

  Derivando con respecto a la tensión se obtiene:

$$C_{gs} = \frac{2}{3} WL C_{ox} + C_{solapamiento} \tag{5.36}$$

  donde $C_{solapamiento}$ es la capacidad asociada al pequeño solapamiento que existe entre el óxido de puerta y la región N^+ de fuente.

La disposición final de todos estos parámetros entre los cuatro terminales del dispositivo se puede ver en la Figura 5.5.4.

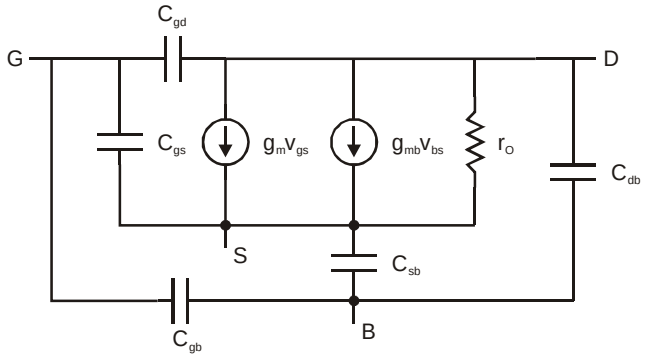


Figura 5.5.4 Modelo de pequeña señal de un MOSFET.

5.6 Uso de los modelos del MOSFET en circuitos

Considérese el circuito de la Figura 5.6.1 en el que aparece una fuente de alimentación V_{DD} y una fuente de pequeña señal v_i. En el apartado anterior se han encontrado modelos lineales para los regímenes de gran y pequeña señal. Este hecho permite que en primer lugar se pueda aplicar el principio de superposición y, posteriormente, analizar por separado la respuesta del circuito a cada una de las fuentes.

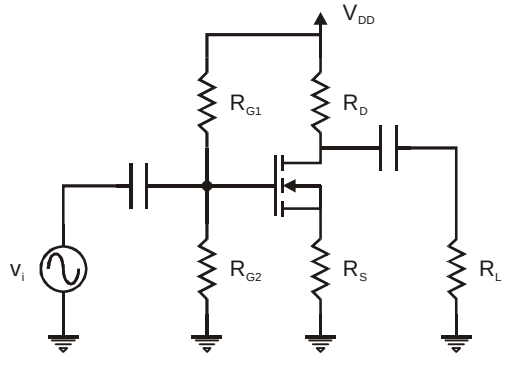


Figura 5.6.1 Circuito real con un MOSFET donde se combinan fuentes de corriente continua y alterna.

Si se estudia el régimen de corriente continua se anula la fuente v_i y los condensadores permanecen en circuito abierto. El circuito se transforma en el que se muestra en la Figura 5.6.2.a. En la Figura 5.6.2.b se ha sustituido el MOSFET por su modelo equivalente. A partir de aquí habría que hacer uso de las leyes de Kirchoff para analizar el circuito.

En el régimen de pequeña señal se anula la fuente de tensión continua. El circuito resultante se muestra en la Figura 5.6.3.a. En la Figura 5.6.3. b se ha eliminado el transistor y en su lugar se debe colocar

el modelo equivalente de pequeña señal. No se ha incorporado en el dibujo para no añadir complejidad al mismo.

En cualquier caso, en la mayoría de las situaciones prácticas no se consideran todos los elementos del modelo. Normalmente se utilizan modelos simplificados, especialmente cuando se hace análisis a mano de circuitos. Cuando se hace necesario un análisis más preciso se utilizan programas de simulación de circuitos. En estos casos sí se incorporan todos los elementos.

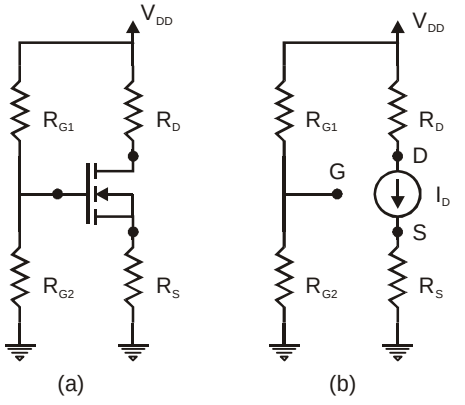


Figura 5.6.2 Respuesta del transistor de la Figura 5.6.1 a la fuente de corriente continua y sustitución del mismo por su modelo equivalente de gran señal.

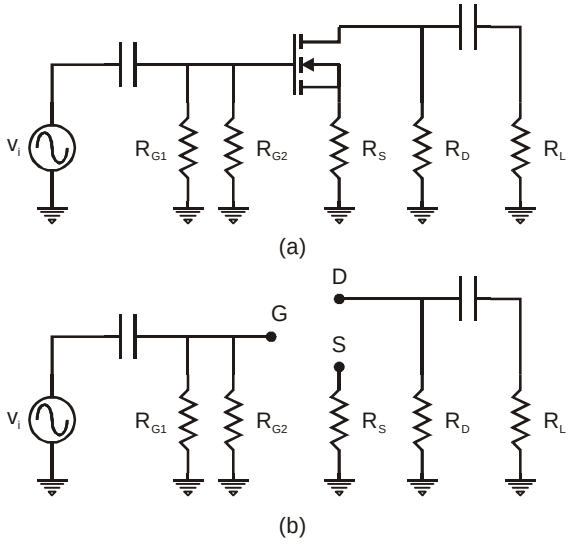


Figura 5.6.3 Respuesta del transistor de la Figura 5.6.1 a la fuente de señal alterna y sustitución del mismo por su modelo equivalente de pequeña señal.

5.7 Respuesta en frecuencia del MOSFET

Considérese un circuito con un MOSFET en el que se quiere amplificar una corriente y recogerla amplificada a la salida en una carga $R_L = 0$. Dicho circuito deberá contar con fuentes de alimentación para polarizarlo adecuadamente. En este ejemplo vamos a mostrar sólo el análisis de pequeña señal de este circuito. Después de anular las fuentes de continua el circuito queda como en la Figura 5.7.1. En él se muestra la fuente de corriente i_g que se quiere amplificar y el cortocircuito de salida donde se quiere medir la corriente i_d. El resto de elementos que aparecen pertenecen al modelo de un transistor MOSFET. De la configuración de este circuito se deduce que el sustrato y la fuente se encuentran cortocircuitados pues no aparecen los elementos g_{mb} y C_{sb}.

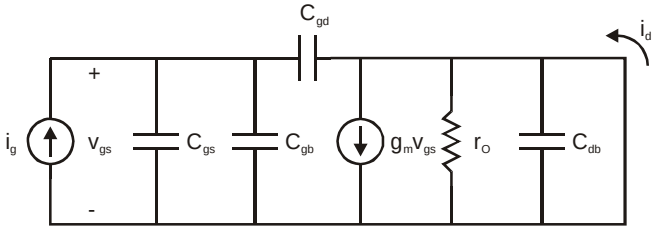


Figura 5.7.1 Respuesta en pequeña señal de un MOSFET excitado con una fuente de corriente y donde la corriente de salida se mide en un cortocircuito.

Para calcular la ganancia en corriente en función de la frecuencia debemos acudir a la notación fasorial y al formalismo de la transformada de Laplace, para lo cual se transforman todos los elementos del circuito y posteriormente se aplican las leyes de Kirchoff. De esta forma las corrientes $I_g(s)$ e $I_d(s)$ quedan de la forma:

$$\begin{aligned} I_g &= V_{gs}(s(C_{gs} + C_{gb} + C_{gd})), \\ I_d &= g_m V_{gs}. \end{aligned} \qquad (5.37)$$

La función de transferencia se obtiene dividiendo las dos corrientes anteriores:

$$H(s) = \frac{I_d}{I_g} = \frac{g_m}{s(C_{gs} + C_{gb} + C_{gd})}. \qquad (5.38)$$

Para conocer la respuesta en frecuencia de este circuito basta evaluar su función de transferencia en $s = j\omega$.

$$H(j\omega) = \frac{g_m}{j\omega(C_{gs} + C_{gb} + C_{gd})}. \qquad (5.39)$$

Se observa como el módulo de la función de transferencia disminuye con la frecuencia. Eso significa que la ganancia de corriente puede tomar valores inferiores a la unidad a partir de una determinada frecuencia. Sería interesante calcular dicha frecuencia pues ella nos permite separar los rangos de frecuencia donde hay amplificación de aquellos en los que hay atenuación. Para ello basta con hacer uno el módulo de la ganancia y despejar la frecuencia:

$$\left|\frac{i_d}{i_g}(j2\pi f_T)\right| = 1 \Rightarrow f_T = \frac{1}{2\pi}\frac{g_m}{C_{gs}+C_{gb}+C_{gd}}. \qquad (5.40)$$

Esta frecuencia se va a denominar frecuencia para ganancia en corriente en cortocircuito unidad, f_T. Vemos que depende de la transconductancia, es decir de la corriente del transistor, y de las capacidades parásitas. Esto demuestra la importancia que tienen estas capacidades a altas frecuencias, pues son las que determinan el decrecimiento de la ganancia a medida que aumenta la frecuencia. En el caso frecuente de que domine la capacidad C_{gs} el parámetro f_T se puede aproximar por:

$$f_T \approx \frac{1}{2\pi}\frac{g_m}{C_{gs}} = 1.5\frac{\mu_n}{2\pi L^2}(V_{GS}-V_t), \qquad (5.41)$$

donde se ha hecho uso de (5.26) y (5.36). Se observa que la f_T depende inversamente del cuadrado de la longitud del canal. Eso significa que si nosotros queremos aumentar el rango de frecuencias de operación de un transistor MOSFET uno de los primero aspectos en los que nos deberíamos fijar sería en reducir las dimensiones del dispositivo. Más adelante justificaremos que en el caso de transistores de longitud de canal inferior a la micra el parámetro f_T es proporcional a $1/L$.

Otra de las conclusiones que se pueden sacar de este estudio es que en el caso de frecuencias bajas o intermedias el efecto de los condensadores del modelo es despreciable. Por ello, se suele emplear un modelo simplificado que permite hacer una estimación rápida del comportamiento de los circuitos con MOSFETs. Este modelo se muestra en la Figura 5.7.2. Este modelo recoge los mismos parámetros que se tuvieron en cuenta cuando se realizó el análisis puramente formal del transistor en el apartado 5.5.2

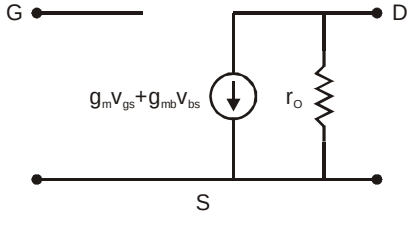


Figura 5.7.2 Modelo simplificado del MOSFET a bajas frecuencias.

5.8 Efectos de canal corto

Se acaba de deducir que la reducción de la longitud del canal comporta beneficios en cuanto que se puede utilizar el transistor a mayores frecuencias. La disminución de dimensiones en esta y las otras direcciones lleva consigo, además de un aumento de la f_T, una reducción de los costes de fabricación y valores más bajos de las capacidades parásitas. En un transistor bipolar la dimensión típica es la anchura de su región de base, medida en sentido vertical al de fabricación del dispositivo, y que puede tomar valores del orden de 0.05 μm. En el caso de un MOSFET su dimensión típica es la longitud del canal, de dirección horizontal, y que toma valores inferiores a 1μm. El uso de transistores de longitud del canal inferior a 1μm obliga a modificar los modelos de gran y pequeña señal del transistor. En los casos que hemos trabajado hasta ahora se ha considerado régimen de bajos campos en el movimiento de los portadores por el canal del MOSFET. Sin embargo, si se reducen las dimensiones del dispositivo pero no se reducen en la misma proporción las tensiones aplicadas entre sus terminales, el campo eléctrico se verá incrementado. Esto tiene una repercusión inmediata: la relación entre la velocidad de los portadores de carga en el canal y el campo eléctrico deja de ser lineal. La Figura 5.8.1 muestra este fenómeno. En ella se representa la velocidad de los electrones en un semiconductor en función del campo eléctrico. Para bajos valores del campo eléctrico la relación velocidad-campo es de la forma:

$$\mathbf{v} = \mu_n \mathbf{E} \,. \tag{5.42}$$

Para altos campos esta relación se puede aproximar empíricamente por:

$$\mathbf{v} \approx \frac{\mu_n \mathbf{E}}{1 + \dfrac{|\mathbf{E}|}{E_c}}, \tag{5.43}$$

donde E_c es el campo crítico, a partir del cual la relación v–E deja de ser lineal. En esta figura $E_c = 1.4\times10^4$ V/cm y la movilidad de los electrones $\mu_n = 700$ cm^{2}V^{-1}s^{-1}.

Para ver cómo afecta esta relación al comportamiento de un MOSFET se puede evaluar la tensión que habría que aplicar al drenador de un MOSFET de canal de 1μm para alcanzar el campo crítico. El resultado sería 1.5 V, por lo que resulta necesario modificar los modelos empleados para describir al dispositivo.

Hasta ahora siempre que se había trabajado con la expresión de la corriente $I = qnvWt$ (donde n es la densidad de electrones, v es su velocidad, y W y t son la anchura y el espesor del canal respectivamente) se habían considerado bajos campos, con lo que se apreciaba un comportamiento óhmico:

$$I = qn\mu_n EWt = \sigma EWt = V / R\,. \tag{5.44}$$

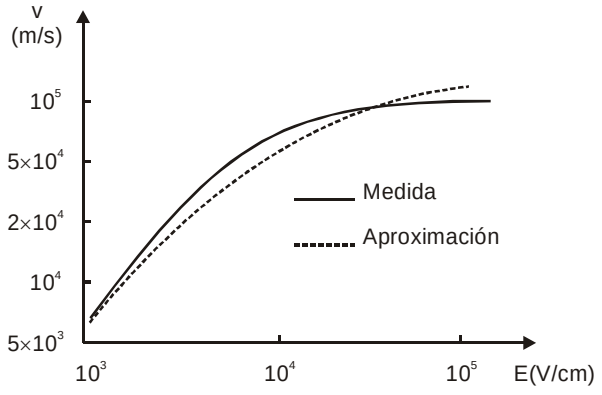


Figura 5.8.1 Velocidad de los electrones en un semiconductor en función del campo eléctrico aplicado al mismo.

De hecho, cuando se calculó la corriente de drenador en el MOSFET, se dividió el canal en elementos de longitud y se trabajó con la resistencia de esos elementos, lo que llevaba implícito el comportamiento óhmico (ecuación (5.6)).

Haciendo uso de la carga por unidad de superficie $Q_I(y) = qnt$ y de la expresión de la velocidad de los electrones a bajos campos (5.42) podemos rescribir la expresión (5.7) correspondiente a la corriente que circula por cada uno de esos elementos:

$$I_D = W Q_I(y) \mu_n \frac{dV}{dy}\,. \tag{5.45}$$

Para altos campos, haciendo uso de (5.43), quedaría de la forma:

$$I_D\left(1 + \frac{1}{E_c}\frac{dV}{dy}\right) = W Q_I(y) \mu_n \frac{dV}{dy}\,. \tag{5.46}$$

Integrando la ecuación anterior a lo largo del canal se obtiene

$$\int_0^L I_D\left(1 + \frac{1}{E_c}\frac{dV}{dy}\right) dy = \int_0^{V_{DS}} W Q_I(y) \mu_n dV\,. \tag{5.47}$$

La solución de esta integral nos proporciona la expresión de la corriente en la región triodo:

$$I_D = \frac{\mu_n C_{ox}}{2\left(1 + \frac{1}{E_c}\frac{V_{DS}}{L}\right)} \frac{W}{L}\left(2(V_{GS} - V_t)V_{DS} - V_{DS}^2\right). \tag{5.48}$$

La corriente en la región de saturación se obtiene evaluando la expresión anterior para $V_{DS} = V_{GS} - V_t$:

$$I_D = \frac{k'}{2\left(1+\theta\left(V_{GS}-V_t\right)\right)}\frac{W}{L}\left(V_{GS}-V_t\right)^2, \qquad (5.49)$$

aunque en la práctica la saturación empieza antes de este valor. En esta expresión se ha definido el parámetro $\theta \equiv 1/(LE_c)$. En este parámetro también se suele incluir la dependencia de la corriente de drenador con el campo perpendicular al canal. Para canales muy cortos la corriente de drenador tiene un comportamiento lineal con $(V_{GS} - V_t)$.

Modelos del MOSFET de canal corto

Al modificar las ecuaciones que describen el comportamiento del MOSFET se deben cambiar en consecuencia los modelos del dispositivo. Considérese un MOSFET de canal largo polarizado en saturación y conectado en serie con el terminal de fuente S' una resistencia de valor R_{SX} (Figura 5.8.2) Al otro extremo de la resistencia se le asigna la letra S.

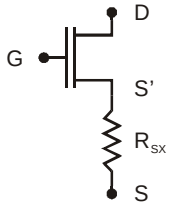


Figura 5.8.2 MOSFET de canal largo con resistencia serie conectada al terminal de fuente.

Se va a considerar todo el circuito como un bloque y se va a extraer una relación entre la corriente I_D con la nuevas tensiones V_{DS} y V_{GS}. Combinando la expresión de la corriente del MOSFET en saturación (5.9):

$$I_D = \frac{\mu C_{ox}}{2}\frac{W}{L}\left(V_{GS'}-V_t\right)^2 \qquad (5.50)$$

y la segunda ley de Kirchoff aplicada a la rama DS:

$$V_{GS} = V_{GS'} + I_D R_{SX}, \qquad (5.51)$$

se obtiene:

$$I_D = \frac{\mu C_{ox}}{2}\frac{W}{L}\left(V_{GS}-I_D R_{SX}-V_t\right)^2, \qquad (5.52)$$

de donde se puede despejar el valor de la corriente I_D:

$$I_D = \frac{\mu C_{ox}}{2\left[1+\mu C_{ox}\frac{W}{L}R_{SX}\left(V_{GS}-V_t\right)\right]}\frac{W}{L}\left(V_{GS}-V_t\right)^2. \qquad (5.53)$$

Se observa que un MOSFET de canal largo con una resistencia en serie con el terminal de fuente proporciona el mismo comportamiento que un MOSFET de canal corto sin más que definir el parámetro θ como:

$$\theta \equiv \mu C_{ox} \frac{W}{L} R_{SX}. \tag{5.54}$$

Si se identifica con el mismo parámetro obtenido anteriormente $\theta \equiv 1/(LE_c)$ se obtiene el valor de la resistencia que modela los efectos de canal corto:

$$R_{SX} = \frac{1}{E_c} \frac{1}{\mu C_{ox}} \frac{1}{W}. \tag{5.55}$$

Para obtener el modelo de pequeña señal habría que recalcular las transconductancias g_m y g_{mb} tomando como punto de partida la nueva expresión para la corriente de drenador (5.53).

5.9 Conducción subumbral en MOSFETs

De acuerdo con los modelos utilizados para estudiar el MOSFET, para que exista corriente entre drenador y fuente debe existir un canal de conducción. Sin embargo, experimentalmente se comprueba que por debajo de la tensión umbral existe una corriente pequeña pero no nula. Una corriente asociada al paso de los electrones que parten de la fuente y deben alcanzar el drenador tras superar la barrera de potencial entre fuente y sustrato. Esa corriente se representa en la Figura 5.9.1.

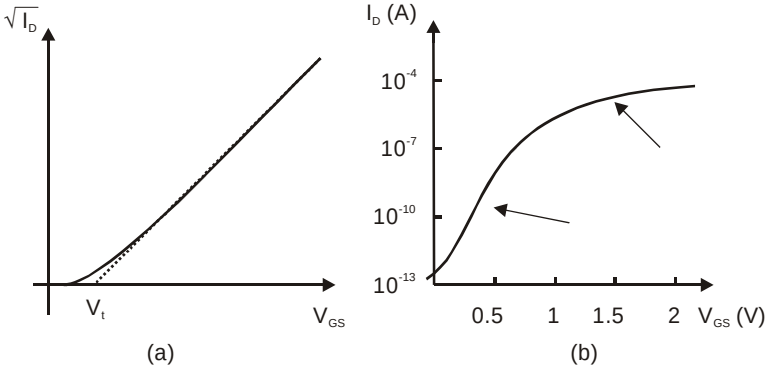


Figura 5.9.1 Corriente de drenador en un MOSFET. Se puede apreciar la corriente no nula para valores inferiores a la tensión umbral.

En la Figura 5.9.1a la relación $I_D^{1/2}$–V_{GS} deja de comportarse como una recta para valores cercanos inferiores a la tensión umbral. Para ver esa región con más detalle se puede hacer una representación logarítmica (Figura 5.9.1b). En ella se observa un cambio de dependencia con la tensión V_{GS}, de cuadrática a exponencial a medida que disminuye

la tensión de puerta. Esa corriente no nula por debajo de la tensión umbral puede justificarse si se observa el diagrama de bandas del MOSFET en esa región de operación (Figura 5.9.2).

La tensión V_{DS} aplicada entre drenador y fuente cae principalmente en la unión de drenador porque entre fuente y sustrato no hay diferencia de potencial (las dos están a tierra). Si se aplicara una tensión negativa al sustrato la barrera en la zona de fuente sería aún mayor. Los electrones que pasan de fuente a drenador deben superar la barrera de fuente. La corriente depende de la altura de esta barrera y es debida a difusión. El término de arrastre es despreciable en el canal. Si tuviéramos un canal más corto las zonas de carga espacial de drenador y fuente se introducirían en el canal curvando las bandas en el mismo (esto puede verse en la Figura 5.9.3 donde se comparan los diagramas de bandas de un transistor de canal largo con uno corto).

En este caso el sustrato ya no tiene sentido como elemento que fija la tensión a cero y por tanto se reduce también la barrera de fuente. El término de arrastre sería más importante aunque está demostrado por diversos autores que sigue siendo despreciable frente al de difusión.

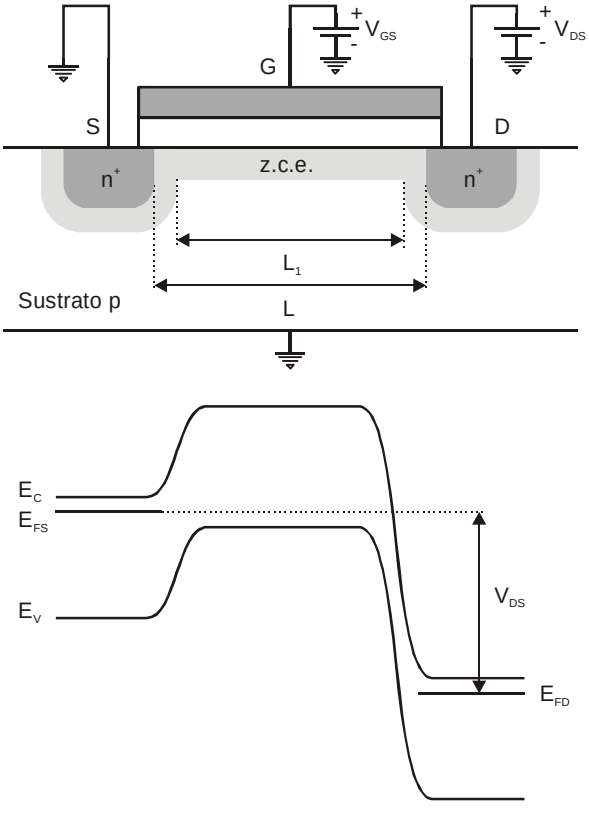


Figura 5.9.2 Diagrama de bandas entre fuente y drenador en un MOSFET de canal largo trabajando en la región subumbral.

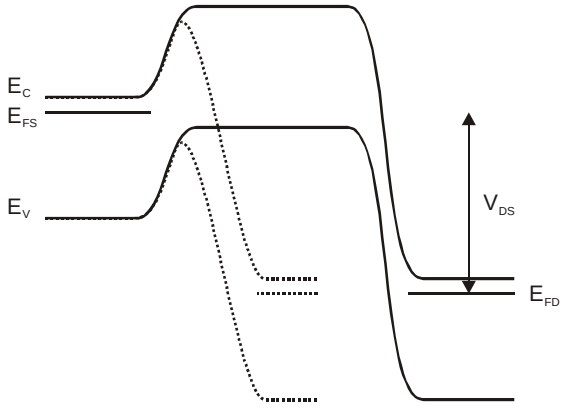


Figura 5.9.3 Comparación de los diagramas de bandas entre drenador y fuente de un MOSFET de canal largo y otro de canal corto.

Por tanto, la corriente subumbral es

$$I_{sub} = -qS_{eff}D_n\frac{dn}{dy}, \tag{5.56}$$

donde S_{eff} es la sección eficaz efectiva para la corriente subumbral $(S_{efff} = Wd_c)$, y d_c es la anchura de la región donde se encuentran la mayoría de los electrones. Si la longitud de difusión de electrones en el sustrato es mayor que la longitud del canal $(L_{nD} >> L)$ la densidad de electrones es lineal, decreciendo de fuente a drenador (no hay recombinación):

$$n(y) = n_{sd} - (n_{sd} - n_{dd})\frac{y}{L_1} \tag{5.57}$$

donde n_{sd} es la concentración de electrones en el canal en el lado de fuente:

$$n_{sd} = n_{p0}e^{\frac{qV(z)}{KT}} \tag{5.58}$$

y n_{dd} es la concentración de electrones en el canal en el extremo de drenador:

$$n_{dd} = n_{p0}e^{\frac{q(V(z)-V_D)}{KT}} \tag{5.59}$$

y $V(z)$ es el potencial eléctrico perpendicular a la superficie. Dicho potencial se puede estimar a partir del diagrama de bandas de la Figura 5.9.4.

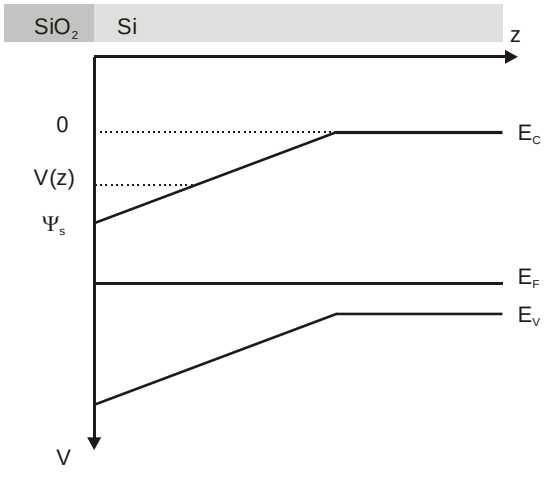


Figura 5.9.4 Diagrama de bandas en el semiconductor en la dirección perpendicular a la superficie Si-SiO_2. Definición del potencial V en un punto del semiconductor separado una distancia z de la superficie.

En dicho diagrama se ha considerado que el campo en la zona de carga espacial es constante e igual a:

$$E_s = \frac{|Q_{dep}|}{\varepsilon_s} = \sqrt{\frac{2q\psi_s N_A}{\varepsilon_s}} \,. \tag{5.60}$$

El potencial en la región de vaciamiento será por tanto lineal:

$$V(z) = \psi_s - E_s z \,. \tag{5.61}$$

De acuerdo con esta expresión la concentración de electrones en la superficie con el óxido es proporcional a $\exp(q\psi_S / KT)$ y decrece como $\exp(qE_S z / KT)$. Podemos considerar por tanto que la mayoría de los electrones se encuentran en una capa de espesor $d_c \approx KT / qE_S$. Con este parámetro se puede calcular finalmente la corriente que circula entre drenador y fuente:

$$I_{sub} = -qS_{eff} D_n \frac{dn}{dy} = -q(W\frac{KT}{qF_y})(\mu_n \frac{KT}{q})\frac{n_{dd} - n_{sd}}{L} =$$

$$= -q(W\frac{KT}{q})\sqrt{\frac{\varepsilon_s}{2q\psi_s N_A}}\mu_n \frac{KT}{q}\frac{n_i^2}{N_A}\frac{\exp\left(\frac{qV(z)}{KT}\right)\left(\exp\left(-\frac{V_D}{KT}\right) - 1\right)}{L}, \tag{5.62}$$

donde se ha admitido que en canales largos $L_1 \approx L$. Haciendo uso de la longitud de Debye:

$$L_{Dp} = \sqrt{\frac{\varepsilon_s KT}{q^2 N_A}} \tag{5.63}$$

y admitiendo que $V(z) \approx \psi_s$ para $0 < z < d_c$ podemos escribir la corriente subumbral:

$$I_{sub} = \varepsilon_s \mu_n \frac{W}{L}\left(\frac{KT}{q}\right)^2 \frac{n_i^2}{N_A^2}\left(\frac{KT}{q\psi_s}\right)^{\frac{1}{2}} \frac{e^{\frac{q\psi_s}{KT}}(1 - e^{-\frac{qV_D}{KT}})}{\sqrt{2}L_{Dp}}. \qquad (5.64)$$

En esta expresión hay todavía una variable interna a la cual no tenemos acceso desde los terminales: el potencial de superficie ψ_s. Necesitamos conocer su relación con la tensión aplicada a la puerta del transistor. Recordemos que esa relación viene descrita por la ecuación (4.5):

$$qV_{GS} = qV_{FB} + q\psi_s - q\frac{Q_s}{C_{ox}}. \qquad (5.65)$$

Introduciendo en esta ecuación la relación entre el campo eléctrico del semiconductor con la carga en el mismo $Q_s = -\varepsilon_s E_s$ se puede escribir:

$$\begin{aligned} \psi_s &= V_{GS} - V_{FB} - \frac{d_i \varepsilon_s}{\varepsilon_i} E_s, \\ \psi_s &= V_{GS} - V_{FB} - \frac{d_i \varepsilon_s}{\varepsilon_i}\sqrt{\frac{2q\psi_s N_A}{\varepsilon_s}}. \end{aligned} \qquad (5.66)$$

Elevando al cuadrado la ecuación anterior se tiene:

$$\psi_s^2 + (V_{FB} - V_{GS})^2 + 2\psi_s(V_{FB} - V_{GS}) = \left(\frac{d_i \varepsilon_s}{\varepsilon_i}\right)^2 \frac{2q\psi_s N_A}{\varepsilon_s} \equiv a_s^2 2\psi_s, \qquad (5.67)$$

donde se ha definido:

$$a_s \equiv \sqrt{qN_A\varepsilon_s}\frac{d_i}{\varepsilon_i}. \qquad (5.68)$$

De (5.67) se puede despejar el valor del potencial de superficie:

$$\begin{aligned} &\psi_s^2 + 2\psi_s(V_{FB} - V_{GS} - a_s^2) + (V_{FB} - V_{GS})^2 = 0, \\ &\psi_s = -(V_{FB} - V_{GS} - a_s^2) \pm \sqrt{(V_{FB} - V_{GS} - a_s^2)^2 - (V_{FB} - V_{GS})^2}, \\ &\psi_s = (V_{GS} - V_{FB} + a_s^2) - a_s\sqrt{a_s^2 + 2(V_{GS} - V_{FB})}. \end{aligned} \qquad (5.69)$$

Teniendo en cuenta la dependencia del potencial de superficie con la tensión de puerta se pueden agrupar constantes y modelar la corriente subumbral por la siguiente expresión:

$$I_D = k_x \frac{W}{L}\exp\left(\frac{qV_{GS}}{nKT}\right)\left(1 - \exp\left(-\frac{qV_{DS}}{KT}\right)\right), \quad n \approx 1.5. \qquad (5.70)$$

Para $V_{DS} > KT/q$, I_{sub} es independiente de V_{DS}. Este resultado era de esperar pues en un canal largo la tensión V_{DS} cae principalmente en la unión de drenador. Se tiene una dependencia exponencial con V_{GS} como se

observa experimentalmente (Figura 5.9.1.b). Desde el punto de vista del dispositivo eso significa que, para valores grandes de V_{DS}, se cumple que $n_{dd} << n_{sd}$, con lo que el gradiente dn/dy no se ve afectado por la tensión V_{DS}.

La aplicación principal del MOSFET operando en esta región es en baja potencia y a frecuencias bajas, ya que se ve obligado por el valor bajo de f_T.

5.10 Corriente en el sustrato de MOSFETs

En el estudio que venimos realizando del MOSFET se ha admitido que la corriente que circula por el sustrato es despreciable pues corresponde a la corriente que atraviesa una unión pn polarizada en inverso. Acabamos de mencionar que la tensión V_{DS} cae principalmente en la unión de drenador. También conocemos que a altas tensiones pueden aparecer mecanismos de generación de pares electrón hueco por avalancha. Los electrones que se generan en estas condiciones cerca de la unión de drenador escapan por el drenador y los huecos hacia el sustrato. Aparece por tanto una corriente parásita en el dispositivo. Dicha corriente se puede modelar con el esquema de la Figura 5.10.1:

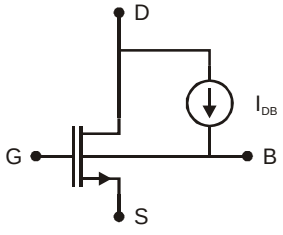


Figura 5.10.1 Modelo del MOSFET incluyendo los efectos de la corriente del sustrato.

donde la fuente de corriente I_{DB} viene descrita por la siguiente ecuación:

$$I_{DB} = k_1 (V_{DS} - V_{DSsat}) I_D \exp\left(-\frac{k_2}{V_{DS} - V_{DSsat}}\right), \qquad (5.71)$$

donde *k1* y *k2* son parámetros que dependen del proceso de fabricación del dispositivo y toman los siguientes valores en transistores NMOS: $k_1 = 5\ \mathrm{V}^{-1}$, $k_2 = 30\ \mathrm{V}$. En transistores de canal P este fenómeno no es tan importante pues los huecos son mucho menos energéticos que los electrones, por lo que son mucho menos eficientes para generar pares electrón hueco.

La repercusión mayor de esta corriente parásita es en el comportamiento de pequeña señal del MOSFET. Este nuevo camino por donde circula la corriente presenta una resistencia que puede ser del orden de otras resistencias del modelo del dispositivo. Este camino paralelo que permite disipar potencia puede ser un factor importante a la hora de

diseñar circuitos con transistores MOS. Para estimar el efecto de dicha corriente se puede calcular la conductancia de pequeña señal asociada, g_{db}. Para ello se deriva la ecuación (5.71):

$$g_{db} = \frac{\partial I_{DB}}{\partial V_{DS}} = k_2 \frac{I_{DB}}{(V_{DS} - V_{DSsat})^2} \cdot \qquad (5.72)$$

Para ver el efecto de este parámetro se puede comparar la inversa de esta conductancia, $r_{db} = 1/g_{db}$, con la resistencia de salida del transistor, r_o. Considérese un transistor caracterizado por los parámetros λ = 0.05 V^{-1} y V_{DSsat} = 0.3 V, por el que circula una corriente I_D = 100 μA y al que se le aplican entre drenador y fuente tensiones de 2 y 4 V. La resistencia de salida de este transistor vale $r_o = 1/(\lambda I_D)$ = 200 KΩ. El valor calculado de r_{db} para estas dos tensiones se encuentra en la tabla:

V_{DS} (V)	r_{db} (Ω)
2	5.3×10^9
4	815×10^3

Resulta evidente que para una tensión de drenador de 2V la conductancia asociada al sustrato es despreciable frente a la conductancia de salida $1/r_o$. Sin embargo, al aumentar la tensión esta conductancia puede resultar comparable a la conductancia de salida. La configuración paralelo equivalente proporciona una conductancia mayor. Este efecto es perjudicial si se pretenden diseñar fuentes de corriente de alta impedancia.

RESUMEN

En este capítulo se ha descrito el transistor de efecto campo metal-aislante-semiconductor. La corriente que circula a través de dichos terminales se ha expresado en función de las tensiones aplicadas al dispositivo.

La característica I-V resultante se ha linealizado para obtener modelos de circuito en gran señal y pequeña señal, mostrando ejemplos de aplicación de dichos modelos. Se han introducido modificaciones a la corriente de drenador debido a efectos de segundo orden como la existencia de altos campos en el canal de conducción, la existencia de una corriente por debajo de la tensión umbral y la corriente que deriva hacia el sustrato.

CUESTIONES Y PROBLEMAS

1. Sea un MOSFET con los siguientes parámetros: Función trabajo metal-semiconductor=-0.1eV, densidad de estados superficiales=+10^{11} átomos/cm^2, espesor del óxido=40 nm, dopado

del sustrato= 10^{16} impurezas/cm^3, profundidad W =2 μm, longitud del canal L =1 μm, V_{SB}=0 V.

a) Represéntese en una gráfica la corriente de drenador en función de la tensión aplicada entre drenador y fuente para los siguientes valores de la tensión V_{GS}: 1.5, 2, 2.5, 3, 3.5 V. Para la zona de inversión considerar dos casos: (i) que está presente el efecto Early y (ii) que es despreciable.

b) Representar en otra gráfica la corriente de drenador en función de V_{GS} en la zona de saturación para los siguientes valores de la tensión V_{SB}: 0, 2, 5, 10, 15 V.

2. Repetir el apartado a) del ejercicio anterior utilizando la expresión de la corriente (5.19) obtenida en el Ejemplo 5.1. En una gráfica aparte comparar la curva obtenida para V_{GS}=3.5V del ejercicio anterior con la que se obtendría con el modelo del citado ejemplo. Incluir una tercera gráfica en la que se haga uso del modelo de canal corto (5.48) y (5.49).

3. Se ha fabricado un MOSFET en un substrato de silicio con una concentración de impurezas de 2×10^{16} cm^{-3}. Se le hace trabajar en saturación aplicándole una diferencia de potencial entre drenador y fuente de 5 V y una tensión puerta fuente V_{GS}. Se mide una corriente de drenador de 12 μA para esta tensión y se obtiene una resistencia de salida de 6 MΩ. Datos tecnológicos: el espesor del óxido es de 400 Å, las dimensiones dibujadas de la puerta del transistor son 7 μm x 100 μm y la difusión lateral de drenador y fuente de 0.6 μm.

a) Averígüese si se trata de un transistor canal N o canal P.

b) Calcular la longitud real del transistor.

c) Calcular la tensión de saturación para esta tensión V_{GS}.

d) Calcular la longitud efectiva del canal.

e) ¿Cuál es el factor de decrecimiento de la longitud del canal al aumentar V_{DS} (dx_d / dV_{DS})?

4. Para un MOSFET de canal N trabajando en la región lineal con V_{DS} = 0.1 V se mide una corriente de 40 μA para V_{GS} = 2 V y 80 μA para V_{GS} = 3 V.

a) ¿Cuál es el valor aparente de la tensión umbral V_t?.

b) Si $k \leftrightharpoons$ 40 μA/V^2 ¿Cuál es el cociente W/L del dispositivo?

c) ¿Qué corriente circulará por el drenador si V_{GS} =2.5 V y V_{DS} = 0.15 V?

d) Si el dispositivo trabaja a V_{GS} =2.5 V ¿para qué valor de V_{DS} se alcanzará el estrangulamiento (pinch-off) en el terminal de drenador? y ¿cuál es la corriente de drenador correspondiente?

5. Sea un transistor MOS de canal N trabajando en la región triodo con valores pequeños de V_{DS} y en un rango de tensiones V_{GS} comprendido entre los 0 V y 5 V. La tecnología de fabricación de estos transistores proporciona un espesor de óxido de 20nm, una longitud de canal superior a 1 μm y una tensión umbral de 0.8 V. ¿Cuál debe ser la profundidad de este dispositivo para conseguir una resistencia de al menos 1 KΩ?

6. Sea un transistor MOSFET de silicio de canal P con las siguientes características: función trabajo metal-semiconductor −0.1 V, densidad de estados superficiales $+10^{11}$ átomos donadores/cm^2, espesor del óxido 400 Å, dopado del sustrato 10^{16} impurezas/cm^3

   a) Calcular la tensión umbral.

   b) Si se implantan impurezas tipo p con una concentración de 9×10^{15} cm^{-3} alcanzando una profundidad de 0.3 μm ¿cuál es la nueva tensión umbral?

   c) Si la profundidad de la implantación fuera de 3 μm ¿cuanto valdría la tensión umbral?

   Sol.: a) -1.5 V, b) -1 V, c) -1.159 V

7. Estimar el tanto por ciento de la corriente de drenador que se pierde por el sustrato en un MOSFET de canal N cuando circula una corriente de drenador de 100 μA y se aplican las siguientes tensiones entre drenador y fuente: V_{DS} =1 V y V_{DS} =5 V. La tensión drenador fuente de saturación es de 0.3 V. Si se necesitan otras constantes usar valores típicos.

   Sol.: 8.5×10^{-17} %, 3.97%

8. Para un transistor MOS de canal N se ha medido una tensión umbral de 0.793 V a temperatura ambiente. Este transistor tiene las siguientes características: concentración de impurezas en el sustrato 5×10^{15} cm^{-3}, densidad de estados superficiales $+10^{11}$ cm^{-2}, función trabajo metal-semiconductor −0.1 V, (V_{SB}=0 V).Si sobre el sustrato de este tipo de transistores se implantan impurezas de fósforo en una concentración 4×10^{15} cm^{-3} hasta una profundidad d, la nueva tensión umbral es de 0.626 V. Estimar la profundidad de la implantación.

   Sol.: 1.04 μm.

9. Calcular el campo eléctrico en el óxido de puerta de un MOSFET de silicio de canal N con espesor t_{ox}=0.1 μm cuando se aplica una tensión de puerta $V_{GS} = 5\ V$ y una tensión de drenador $V_{DS} = 4\ V$, en los siguientes casos:

   a) En un punto próximo a la fuente (x = 0) y en otro próximo al drenador. Hacer uso de las gráficas de la **Figura P.1** y de los siguientes datos: densidad de estados superficiales: $+10^{11}$

cm^{-2}, densidad de carga en el óxido: 1.6×10^{-8} C/cm^2. ¡¡Averiguar el signo de la carga en el semiconductor!!.

b) A partir uso de las gráficas determinar la tensión de banda plana y delimitar sobre las figuras las regiones de acumulación, deplexión e inversión. Continuar la representación del módulo de la carga en el semiconductor para valores negativos del potencial de superficie.

c) Comparar el resultado con el campo eléctrico necesario para la avalancha en una unión pn (3×10^5 V/cm para dopados 10^{15}–10^{16} cm^{-3} y 10^6 V/cm para dopados 10^{18} cm^{-3}.

d) Llamemos *EMAX* al máximo del campo obtenido en los apartados a) y c). Si tenemos una unión pn abrupta con concentraciones uniformes de impurezas donadoras (10^{16} cm^{-3} en el lado n) y de impurezas aceptadoras (10^{15} cm^{-3} en el lado p), ¿qué tensión (*VMAX*) deberíamos aplicar a los extremos de la unión para que el campo máximo en la misma coincidiera con *EMAX*?.

e) Si la unión anterior se polariza en inverso con una fuente de tensión de valor *VMAX* + 10 V en serie con una resistencia de 10 KΩ, ¿qué corriente circula por el circuito? AYUDA: representar la característica I-V en inverso determinando bien cual es la tensión de ruptura.

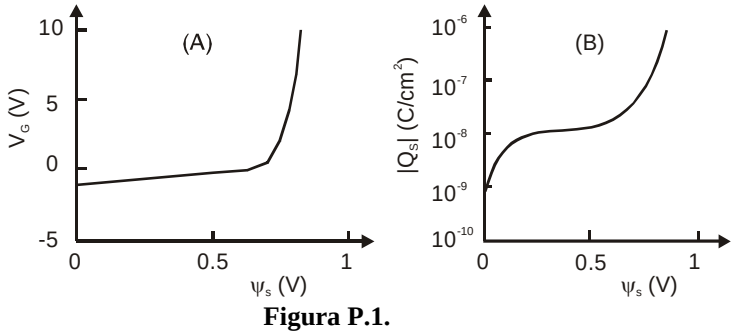


Figura P.1.

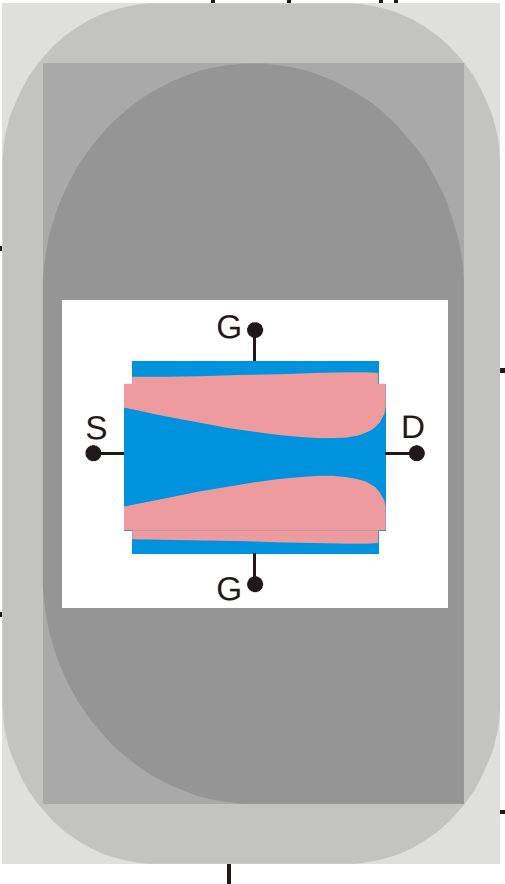
REFERENCIAS

[1] Y. P. Tsividis, *Operation and modelling of the MOS Transistor*, McGraw Hill, 1987.

[2] P.R.Gray and R.G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 3ª Ed., John Wiley & Sons, 1993.

Capítulo 6

El transistor de Efecto Campo de Unión (JFET).



Transistor Jfet

Índice

Objetivos

- Definir las regiones que componen el transistor de efecto campo de unión.
- Describir el funcionamiento básico de este dispositivo.
- Evaluar de forma cualitativa la corriente que va a circular por el canal del transistor en función de las tensiones aplicadas a sus terminales.
- Presentar un modelo que permita encontrar expresiones analíticas para la corriente que circula por el transistor.
- Proporcionar modelos eléctricos equivalentes en gran señal y pequeña señal. Los modelos deberán ser lineales y permitirán analizar o diseñar circuitos electrónicos que incluyan a este dispositivo.
- Simplificar los modelos equivalentes encontrados para facilitar el análisis de estos circuitos.

Palabras Clave

Transistor de efecto campo de unión.
Dispositivo unipolar.
Espesor de la zona de carga espacial.
Espesor del canal de conducción.
Tensión de agotamiento.
Corriente de drenador.
Tensión de puerta.
Tensión de drenador.
Región triodo.
Región de saturación.
Tensión de saturación.
Modulación de la longitud del canal.
Tensión Early.
Región de ruptura.
Modelo analítico de uniones abruptas.
Modelo de gran señal.
Modelo de pequeña señal.
Transconductancia.
Resistencia de salida.
Capacidades de las uniones.
Resistencias parásitas.
Respuesta en frecuencia.

6.1 Estructura y funcionamiento

Transistor de efecto campo de unión:
La corriente que circula entre fuente y drenador se controla modificando el espesor del canal. Esto se consigue variando la zona de carga espacial de las dos uniones que constituyen el dispositivo. Dichas uniones se polarizan en inverso mediante el terminal de puerta.

Como transistor de efecto campo la idea es controlar la corriente que circula por un canal semiconductor mediante un campo eléctrico aplicado a un terminal de puerta. A diferencia del MOSFET, la estructura que nos encontramos ahora en la puerta es una unión polarizada en inverso (para evitar que haya corriente a través de ella). El diseño de dispositivo se muestra en la Figura 6.1.1.

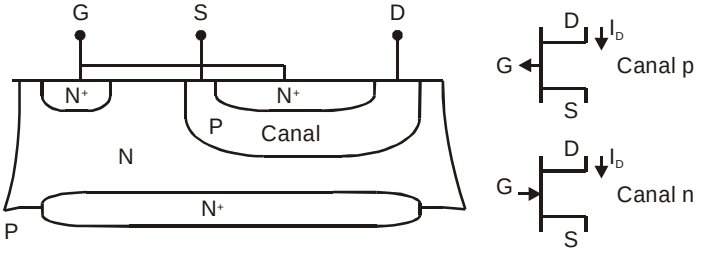


Figura 6.1.1 Estructura de un JFET de canal P y símbolos para el transistor de canal P y N.

En ella vemos un semiconductor tipo P en cuyos extremos conectamos los terminales de fuente y drenador. A su vez se encuentra entre dos semiconductores tipo N sobre los que se conecta el terminal de puerta. La región N^+ que se encuentra enterrada en la parte inferior del dispositivo sirve para minimizar resistencias parásitas. Estas resistencias están asociadas a los caminos que debe seguir la corriente hasta llegar a la zona activa del dispositivo. La aplicación de una diferencia de potencial entre fuente y drenador da lugar a un flujo de huecos (portadores mayoritarios en el canal) entre estos dos terminales. Es un dispositivo unipolar, al igual que el MOSFET, pues no participan portadores minoritarios en la corriente.

Podemos encontrar igualmente transistores de canal N, es decir, un semiconductor N entre otros dos tipo P. En este caso la conducción a través del canal es debida a electrones. Para distinguir estos dos tipos de transistores se utilizan los símbolos de la Figura 6.1.1.

Modo de operación

El efecto de la puerta como terminal de control se puede ver en la Figura 6.1.2. Al aumentar la tensión de puerta aumenta la zona de carga espacial disminuyendo la sección del canal. Esto permite controlar la resistencia del semiconductor y por tanto la corriente que por él circula.

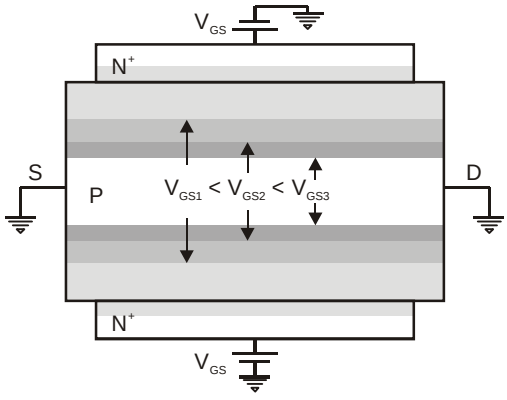


Figura 6.1.2 Control del espesor del canal variando la tensión aplicada a la puerta.

La zona de carga espacial se extiende y varía principalmente en la región menos dopada. A diferencia del MOSFET aquí existe canal Para V_{GS} =0. Sin embargo, se alcanzará una tensión a la cual la zona de carga espacial ocupe todo el canal. A la tensión que origine este fenómeno se denomina tensión de agotamiento o "pinch-off" (V_{GS} =V_p).

Tensión de agotamiento V_p : Diferencia de potencial aplicada entre la puerta y un punto del canal semiconductor a la cual la zona de carga espacial de las dos uniones en esa zona del canal se hace igual al espesor del canal.

Su cálculo es sencillo pues basta igualar la anchura de la zona de carga espacial de la región P de una de las uniones (admitiéndolas iguales) a la mitad de la anchura del canal:

$$a = \sqrt{\frac{2\,\varepsilon_s(\psi_0 + V_p)}{q\,N_A(1+\frac{N_A}{N_D})}}\,, \qquad (6.1)$$

donde N_A y N_D son las concentraciones de impurezas en la zona P y N respectivamente y Ψ_0 es el potencial barrera de la unión. Despejando el valor de la tensión de agotamiento se tiene:

$$V_p = a^2 \frac{q\,N_A(1+\frac{N_A}{N_D})}{2\,\varepsilon_s} - \psi_0\,. \qquad (6.2)$$

La tensión de agotamiento suele tomar valores entre 1 y 3 V. Presenta una dependencia térmica del orden de –2mV/ºC.

6.2 Cálculo cualitativo de la característica corriente-tensión

La relación entre la corriente de drenador y las tensiones V_{DS} y V_{GS} en el JFET es muy similar a la de otro transistor de efecto campo. Basta con analizar la evolución de la geometría del canal que se muestra en la Figura 6.2.1.

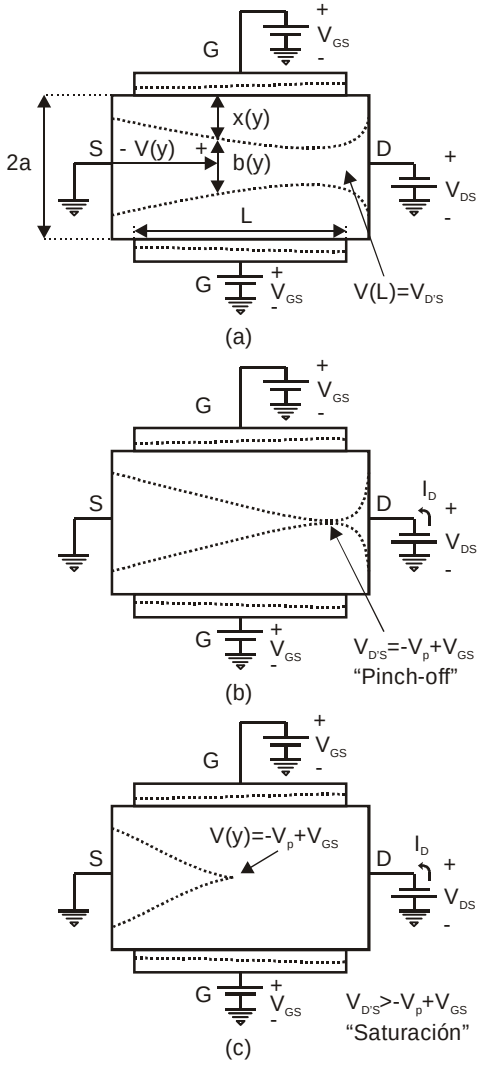


Figura 6.2.1 Evolución del canal a medida que se incrementa la magnitud de la tensión V_{DS}, para una tensión V_{GS} constante, en un JFET de canal P.

En esas figuras se mantiene la tensión V_{GS} a un valor constante y se aumenta el valor de la fuente de drenador. En ellas se muestra como la zona de drenador es diferente de la de fuente pues la diferencia de

potencial entre la zona N de puerta y la P de canal varía con la posición, alcanzando un valor mayor en el drenador. El canal en esta zona es por tanto más estrecho. La manera de cambiar la geometría del canal es muy similar a la del canal de un MOSFET. Por consiguiente, es esperable que se obtengan unas curvas también similares, donde se distinga una región triodo, con un comportamiento lineal entre I_D y V_{DS} para valores bajos de la tensión de drenador, y termine la corriente por saturarse a un valor aproximadamente constante (no es estrictamente constante pues, al igual que en el MOSFET, también aparece el efecto Early).

En la Figura 6.2.2 se muestra una colección de curvas para un transistor de canal P. Se ha incluido también la región de ruptura, que refleja los mecanismos de generación por avalancha en la unión drenador-puerta para altos valores de la tensión V_{DS}. Se observa como a medida que aumenta la tensión de puerta la ruptura se alcanza para un valor menor de V_{DS}. Ello es lógico pues el aumento de tensión en la puerta de debe compensar con una disminución del potencial en el drenador. El parámetro BV_{DG0} se define como la tensión de ruptura en la unión drenador-puerta con la fuente abierta. El signo negativo del potencial V_{DS} y de la corriente I_D resultan de la definición de variables que se halla en la Figura 6.2.1. Se observa como la magnitud de la corriente disminuye a medida que aumenta la tensión de puerta.

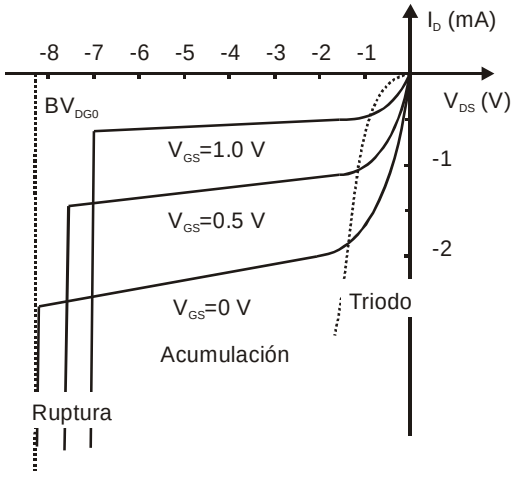


Figura 6.2.2 Curvas características en un JFET de canal P. Definición de las regiones triodo, saturación y ruptura.

6.3 Modelo de uniones abruptas para el cálculo de la característica corriente tensión

Consideremos el transistor que se muestra en la Figura 6.3.1.

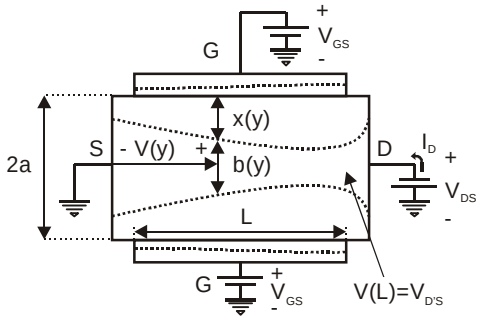


Figura 6.3.1 Definición de las magnitudes que intervienen en el cálculo de la característica corriente tensión en el JFETde canal P.

Sean *y* un punto del canal tomando la fuente como origen, *L* la longitud del canal, *V(y)* el potencial en ese punto *y*, *x(y)* la anchura de la zona de carga espacial en la zona P y *b(y)* la anchura del canal en ese punto. La anchura de la zona de carga espacial *x(y)* se puede relacionar con la diferencia de potencial entre la tensión puerta y la tensión del punto *y*, $V_R(y) \equiv V_{GS} - (y)$:

$$\begin{aligned} x(y) &= k_1\sqrt{\psi_0 + V_R(y)}, \\ x(y) &= k_1\sqrt{\psi_0 + V_{GS} - V(y)}, \end{aligned} \tag{6.3}$$

donde se ha definido

$$k_1 = \sqrt{\frac{2\varepsilon_s}{q N_A (1 + \frac{N_A}{N_D})}}\,. \tag{6.4}$$

La anchura del canal se puede relacionar con *x(y)*:

$$b(y) = 2a - 2x(y) = 2a - 2k_1\sqrt{\psi_0 + V_{GS} - V(y)}\,. \tag{6.5}$$

La densidad de corriente en el punto *y* se puede expresar en función del campo eléctrico:

$$\begin{aligned} J_y &= \sigma E_y, \\ -\frac{I_D}{Wb(y)} &= \sigma\frac{-dV(y)}{dy}. \end{aligned} \tag{6.6}$$

Integrando esta ecuación a lo largo de todo el canal se obtiene:

$$I_D\int_0^L dy = \sigma W\int_0^{V(L)} b\,dV\,. \tag{6.7}$$

Definiendo $V_{D'S} \equiv V(L)$ y teniendo en cuenta que $|V_{DS}| > |V(L)|$ el resultado de la integral es:

$$I_D = G_0\left[V_{D'S} + \frac{2}{3}\frac{k_1}{a}(\psi_0 + V_{GS} - V_{D'S})^{\frac{3}{2}} - \frac{2}{3}\frac{k_1}{a}(\psi_0 + V_{GS})^{\frac{3}{2}}\right], \qquad (6.8)$$

donde se ha definido $G_0 = 2a\sigma W / L$. Haciendo uso de la condición de agotamiento del canal:

$$a = k_1\sqrt{\psi_0 + V_p} \qquad (6.9)$$

se puede escribir:

$$I_D = G_0\left[V_{D'S} + \frac{2}{3}\frac{(\psi_0 + V_{GS} - V_{D'S})^{\frac{3}{2}} - (\psi_0 + V_{GS})^{\frac{3}{2}}}{(\psi_0 + V_p)^{\frac{1}{2}}}\right]. \qquad (6.10)$$

Esta expresión corresponde a la región triodo. Es válida para valores de tensión inferiores a la condición de agotamiento $|V_{D'S}| < |V_{Dsat}| \equiv |V_{GS} - V_p|$. En la Figura 6.3.2 podemos ver una serie de curvas que corresponden a distintos valores de la tensión V_{GS} y el valor de $V_{D'S}$ donde termina la región triodo y comienza la región saturación.

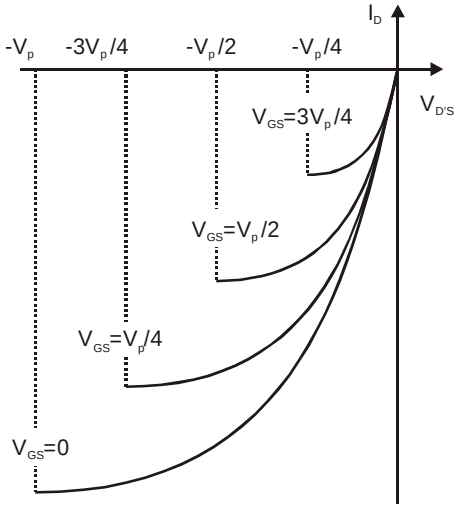


Figura 6.3.2 Representación de curvas de corriente en la región triodo junto con los valores de la tensión $V_{D'S}$ que delimitan dicha región con la de región de saturación.

Para encontrar la expresión de la corriente en la región de saturación basta con evaluar la ecuación (6.10) para $V_{D'S} = V_{GS} - V_p$:

$$I_D = I_D(V_{D'S} = -V_p + V_{GS}), \qquad (6.11)$$

$$I_D = G_0\left[-V_p + V_{GS} + \frac{2}{3}\frac{(\psi_0 + V_p)^{\frac{3}{2}} - (\psi_0 + V_{GS})^{\frac{3}{2}}}{(\psi_0 + V_p)^{\frac{1}{2}}}\right]. \qquad (6.12)$$

Esta expresión se maximiza para V_{GS}=0 y toma el valor:

$$I_{DSS} \equiv I_D(V_{GS} = 0) = G_0\left[-V_p + \frac{2}{3}\frac{(\psi_0 + V_p)^{\frac{3}{2}} - (\psi_0)^{\frac{3}{2}}}{(\psi_0 + V_p)^{\frac{1}{2}}}\right]. \qquad (6.13)$$

Normalmente se suele utilizar una expresión empírica más compacta para la zona de saturación, en la que aparece una dependencia cuadrática con la tensión V_{GS}:

$$I_D \approx I_{DSS}\left(1 - \frac{V_{GS}}{V_p}\right)^2. \qquad (6.14)$$

Para incluir el efecto de la modulación de la longitud del canal se añade el parámetro λ cuyo inverso nos proporciona la tensión Early.

$$I_D = I_{DSS}\left(1 - \frac{V_{GS}}{V_p}\right)^2 (1 + \lambda V_{DS}). \qquad (6.15)$$

Ejemplo 6.1

Considerar un JFET de canal N con los siguientes parámetros: N_D =10^{16} cm^{-3}, W/L=10, N_A =10^{18} cm^{-3}, L=10 μm, a=1 μm. Representar la relación I_D-V_{GS} en saturación obtenida con el modelo de control de carga (ecuación (6.12)). Compararla con la relación cuadrática (6.14) en la misma gráfica.

Solución.

Para evaluar la expresión de la corriente de drenador es necesario calcular en primer lugar el potencial barrera de la unión y la tensión de agotamiento (6.2):

$$\psi_0 = 0.855\,V$$
$$-V_p = -6.111\,V$$

A partir de esos valores se puede calcular el parámetro I_{DSS}:

$$I_{DSS} = 0.08\,mA$$

El signo positivo de la corriente significa que el sentido real de la corriente es de drenador a fuente, al revés que en un transistor canal P. La tensión puerta fuente también cambia de signo para que las uniones sigan en inversa, por ello la tensión umbral es negativa. Con estos valores se

pueden representar las dos funciones. La Figura 6.3.3 nos muestra el grado de aproximación entre ambas expresiones.

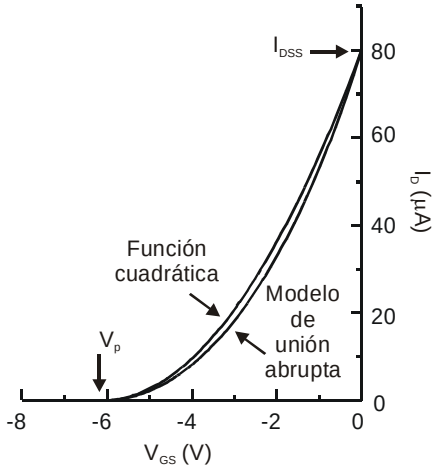


Figura 6.3.3 Corriente de drenador en saturación.

6.4 Modelos de gran señal y pequeña señal en saturación

Conocemos el valor de las corrientes que circulan por todos los terminales del transistor en función de las diferencias de potencias existentes entre ellos: se ha obtenido una expresión para I_D y sabemos que la corriente que circula por la puerta corresponde a la de una unión polarizada en inverso, del orden de 10^{-12}-10^{-10} A. Se puede por tanto modelar al transistor. En la región de saturación el modelo de gran señal del transistor se muestra en la Figura 6.4.1. Se puede ver la similitud entre este modelo y el del MOSFET. La única diferencia es el valor de la fuente dependiente de corriente.

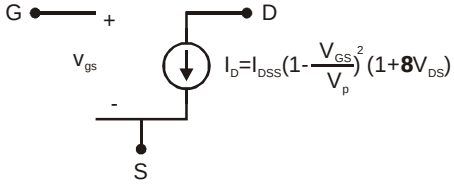


Figura 6.4.1 Modelo del gran señal del JFET en saturación.

Para obtener el modelo de pequeña señal en saturación se puede actuar de la misma manera que con el MOSFET. Tendremos los siguientes elementos:

- Transconductancia g_m:

$$g_m = \frac{dI_D}{dV_{GS}} = -\frac{2I_{DSS}}{V_p}\left(1-\frac{V_{GS}}{V_p}\right) = g_{m0}\left(1-\frac{V_{GS}}{V_p}\right). \qquad (6.16)$$

Con valores típicos de I_{DSS}=-1 mA y V_p =2 V encontramos una transconductancia g_{m0}=1 mA/V.

- Resistencia de salida r_o:

$$\frac{1}{r_o} = \frac{\partial I_D}{\partial V_{DS}} = \lambda I_{DSS}(1 - \frac{V_{GS}}{V_p})^2 \approx \lambda I_D \,. \tag{6.17}$$

- Capacidades asociadas a las zonas de carga espacial de las uniones:
  - ✓ Capacidad puerta-fuente:

$$\frac{C_{gs0}}{(1 + \frac{V_{GS}}{\psi_0})^{\frac{1}{3}}} \,. \tag{6.18}$$

  - ✓ Capacidad puerta-drenador:

$$C_{gd} = \frac{C_{gd0}}{(1 + \frac{V_{GD}}{\psi_0})^{\frac{1}{3}}} \,. \tag{6.19}$$

  - ✓ Capacidad puerta-sustrato.

$$C_{gss} = \frac{C_{gss0}}{(1 + \frac{V_{GSS}}{\psi_0})^{\frac{1}{2}}} \,. \tag{6.20}$$

donde V_{GS}, V_{GD} y V_{GSS} son la tensiones de las uniones respectivas y C_{gs0}, C_{gd0} y C_{gss0} son las capacidades de las uniones a tensión cero.

- Resistencias parásitas de fuente y drenador: r_s y r_d. Está asociadas a los caminos resistivos por los que circula la corriente y unen los contactos con la zona activa del canal. El efecto de la resistencia de fuente repercute en que I_{DSS} y g_m toman valores más pequeños. Al incluirse su efecto en estos parámetros no se incluye explícitamente en el modelo, que toma la forma definitiva que se muestra en la Figura 6.4.2.

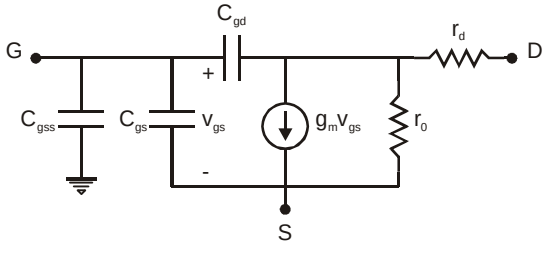


Figura 6.4.2 Modelo de pequeña señal del JFET.

Al igual que se estudió la respuesta en frecuencia del MOSFET se puede hacer un análisis similar con el JFET. Se puede definir

igualmente una frecuencia de ganancia en corriente en cortocircuito unidad, f_T. Se encontraría una expresión como

$$f_T = \frac{1}{2\pi}\frac{g_m}{C_{gs}+C_{dg}+C_{gss}} \tag{6.21}$$

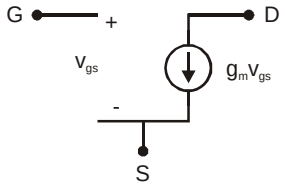


Figura 6.4.3 Modelo de pequeña señal simplificado válido para baja frecuencia.

Introduciendo valores típicos para g_m=1 mA/V y C_{gs}+C_{dg}+C_{gss}=5 pF se obtendría una f_T=30 MHz.

A frecuencias bajas e intermedias el modelo de pequeña señal se puede simplificar por el que aparece en la Figura 6.4.3.

Ejemplo 6.2

Deducir una expresión para la conductancia del canal de un JFET de canal P operando en la región lineal, es decir, para valores bajos de la tensión aplicada entre drenador y fuente, manteniendo fija la tensión V_{GS}.

Solución.

Otra forma de expresar la ecuación (6.10) es:

$$I_D = I_p\left[3\frac{V_{D'S}}{V_a}+2\frac{(\psi_0+V_{GS}-V_{D'S})^{\frac{3}{2}}-(\psi_0+V_{GS})^{\frac{3}{2}}}{V_a^{3/2}}\right], \tag{6.22}$$

donde se ha definido $V_a \equiv \psi_0 + V_p$ e $I_p \equiv V_a G_o / 3$. La ecuación (6.22) se puede rescribir:

$$I_D = I_p\left[3\frac{V_{D'S}}{V_a}+2\frac{(\psi_0+V_{GS})^{\frac{3}{2}}}{V_a^{3/2}}\left\{\left(1-\frac{V_{D'S}}{\psi_0+V_{GS}}\right)^{3/2}-1\right\}\right]. \tag{6.23}$$

Si $V_{D'S} << \psi_0 + V_{GS}$ podemos aproximar la corriente por:

$$I_D \approx I_p\left[3\frac{V_{D'S}}{V_a}+\frac{(\psi_0+V_{GS})^{\frac{3}{2}}}{V_a^{3/2}}3\frac{V_{D'S}}{\psi_0+V_{GS}}\right] \tag{6.24}$$

La corriente para bajos valores de la tensión de drenador quedaría:

$$I_D = \frac{3I_p}{V_a}\left[1+\left(\frac{\psi_0 + V_{GS}}{V_a}\right)^{1/2}\right]V_{D'S}\,. \qquad (6.25)$$

La conductancia se obtiene derivando la expresión anterior con respecto a $V_{D'S}$:

$$g_D = \frac{\partial I_D}{\partial V_{D'S}} = \frac{3I_p}{V_a}\left[1+\left(\frac{\psi_0 + V_{GS}}{V_a}\right)^{1/2}\right]. \qquad (6.26)$$

RESUMEN

En este capítulo se ha estudiado otro transistor de efecto campo. Se han descrito las partes constituyentes de su estructura, cuál es su funcionamiento básico. Se ha hecho uso de un modelo, el de uniones abruptas y canal gradual, con el que se ha calculado la corriente que circula por el canal del transistor. En el conjunto de ecuaciones que describen la corriente de drenador se encuentra la dependencia explícita con las tensiones aplicadas a los terminales de la estructura. Al igual que con otros dispositivos la característica corriente tensión es no lineal por lo que se han buscado modelos que linealicen esta característica tanto en condiciones de gran señal como de pequeña señal. Estos modelos constituyen la etapa básica para poder analizar a este dispositivo en el seno de un circuito electrónico.

CUESTIONES Y PROBLEMAS

1. Para un JFET canal N de uniones abruptas y concentraciones de impurezas uniformes deducir una expresión para la corriente en la región triodo. Comprobar que se obtiene:

$$I_D = G_0\left[V_{D'S} - \frac{2}{3}\frac{(\psi_0 - V_{GS} + V_{D'S})^{\frac{3}{2}} - (\psi_0 - V_{GS})^{\frac{3}{2}}}{(\psi_0 + V_p)^{\frac{1}{2}}}\right]. \qquad (6.27)$$

   Ahora, la diferencia de potencial entre puerta y canal necesaria para agotarlo es $-V_p$. Encontrar también la expresión para la corriente en saturación.

2. A partir de la conductancia obtenida en el Ejemplo 6.2, para un JFET de canal P trabajando en la región lineal, deducir una expresión para la variación más importante de la conductancia del canal con respecto a la temperatura. Para ello, seguir

considerando que la unión puerta-canal es abrupta y que la movilidad de los mayoritarios en el canal varía como $T^{-3/2}$. Comprobar que se obtiene la siguiente expresión:

$$\frac{1}{g_D}\frac{\partial g_D}{\partial T} = \left\{ -\frac{3}{2} + \frac{\left(E_G + 3KT\right)/q - \psi_0}{2V_a^{1/2}(V_{GS}+\psi_0)^{1/2}\left[1-\left((V_{GS}+\psi_0)/V_a\right)^{1/2}\right]} \right\} \frac{1}{T}. \quad (6.28)$$

3. Para el JFET (de canal N) de la figura calcule el punto de operación, la tensión umbral y comprobar que la tensión total a la salida vale:

$$v_o(t) = 14.7 + 2.4\times10^{-6}\cos(\omega t + \pi) \quad V .$$

Datos: $N_D = 10^{16}$ cm^{-3}, $N_A = 10^{18}$ cm^{-3}, L = µm, $W/L = 10$, a = 1.0 µm, $2a$ = ancho del canal, $v_i(t) = 0.1\cos(\omega t)$ mV, $C\omega \to \infty$, R_1 = 83 KΩ, R_2 = 100 KΩ, R_3 = 8 KΩ, R_4 = 282 KΩ.

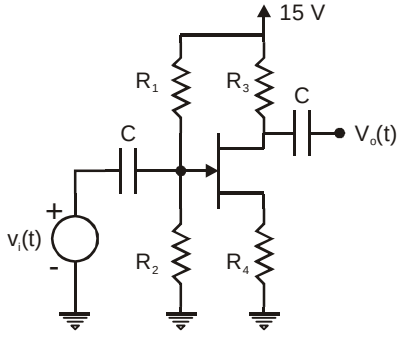


Figura P.1.

4. Se pretende utilizar el JFET de la Figura P.2 como resistencia variable en el intervalo 21.3 kΩ – 26.7 kΩ para valores de la tensión de puerta entre 10 V y 20 V. Suponiendo pequeños valores de la tensión V_{DS} calcular cual debe ser la longitud del canal y la concentración de impurezas del mismo conocidos el resto de los parámetros que se muestran en la figura.

(Sol.: $N_A = 1.27\times10^{17}$ cm^{-3}, L = 266 µm).

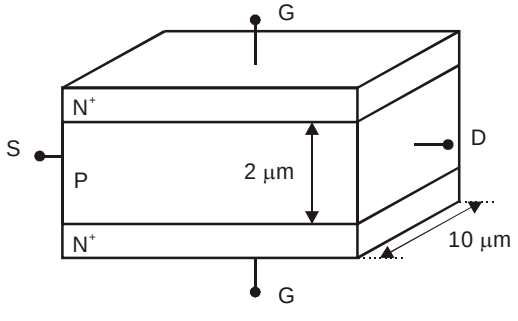


Figura P.2.

REFERENCIAS

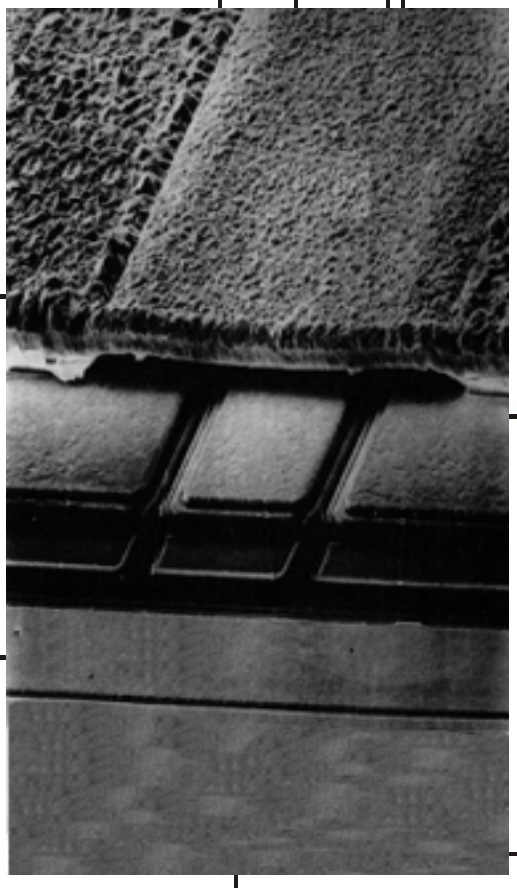
[1] P.R.Gray and R.G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 3ª Ed., John Wiley & Sons, 1993.

[2] S. M. Sze, *Physics of semiconductor devices*, John Wiley & Sons, 1981.

Capítulo 7

El transistor de Efecto Campo Metal Semiconductor

Transistor MESFET



Índice

Objetivos

- Presentar otros materiales semiconductores empleados en aplicaciones electrónicas particulares.
- Describir la estructura del MESFET y su modo de operación.
- Presentar distintos modelos que tienen como fin el llegar a una expresión analítica para la corriente de drenador.
- Presentar otro modelo empleado en diseño asistido por ordenador que proporciona una relación I-V válida para cualquier valor de la tensión.
- Analizar cómo se modifican las expresiones de la corriente cuando se tienen en cuenta otros efectos como las resistencias parásitas o la corriente de puerta a través del contacto Schottky.
- Proponer un método para extraer los parámetros que aparecen en las características I-V a partir de medidas experimentales.
- Presentar un modelo de pequeña señal válido para muy altas frecuencias, rango donde es especialmente útil este dispositivo.
- Mostrar ejemplos de aplicación de este modelo.

Palabras Clave

GaAs.
Contacto Schottky de puerta.
Zona de carga espacial.
Tensión de agotamiento.
Conductividad del canal.
Modelo de canal gradual.
Modelo de dos tramos para la velocidad de los electrones.
Modelo de Curtice.
Resistencias parásitas.
Corriente de puerta.
Extracción de parámetros.
Modelo de pequeña señal.

7.1 El arseniuro de galio (GaAs)

La mayoría de los transistores de efecto campo están fabricados de silicio por las excelentes propiedades de este material y por su abundancia en la naturaleza. Sin embargo, también se utilizan materiales compuestos en su fabricación para utilizarlos en ciertas aplicaciones como alta velocidad, alta frecuencia, o en situaciones donde se someten a los circuitos a condiciones de operación extremas, como alta y bajas temperaturas y exposición a la radiación.

La tecnología de materiales compuestos está menos desarrollada que la del silicio por lo que se prefiere este material en la mayoría de las aplicaciones. Sin embargo, el GaAs y otros materiales compuestos presentan ciertas ventajas sobre el silicio. Entre las ventajas hay que citar (a) que es un material de ancho de banda prohibida directo, con lo que es preferible en aplicaciones optoelectrónicas, (b) los electrones en este material tienen mayor movilidad, con lo que se obtienen resistencias parásitas menores, proporciona una mayor velocidad al funcionamiento del dispositivo y un aumento de las frecuencias de operación, (c) la posibilidad de utilizar sustratos semiaislantes permite a este material usarlo como base de los circuitos integrados monolíticos de microondas y ondas milimétricas. Entre las desventajas que presenta este material frente al silicio hay que mencionar (a) una conductividad térmica menor, con los problemas que esto conlleva a la hora de eliminar la potencia disipada en los circuitos, (b) ausencia de unóxido de calidad y (c) los mayores costes de producción.

7.2 Estructura del MESFET

Transistor de efecto campo metal semiconductor:
La corriente que circula entre fuente y drenador se controla modificando el espesor del canal. A semejanza del JFET, esto se consigue variando la zona de carga espacial en un semiconductor. La diferencia con el JFET estriba en que, en lugar de uniones pn, la puerta del transistor la constituye una unión entre un metal y un semiconductor.

El transistor de efecto campo basado en materiales compuestos más importante es el transistor de efecto campo metal semiconductor (MESFET). El MESFET de GaAs (Figura 7.2.1) consta de un sustrato semiaislante o ligeramente dopado tipo P. Encima de este semiconductor hay otra capa conductora tipo N (canal) en la que se conectan tres terminales: los contactos óhmicos de fuente y drenador y un contacto Schottky que actúa como puerta. Se pueden encontrar también canales tipo P, sin embargo, son menos frecuentes puesto que la movilidad de los huecos es inferior a la de los electrones. Entre fuente y drenador se hará circular una corriente mediante una tensión aplicada entre estos terminales. Esta corriente se puede controlar a su vez mediante una tensión aplicada al terminal de puerta. El control de esta corriente se realiza gracias a la variación de la zona de carga espacial que aparece en el canal debajo de la puerta. Esta región de vaciamiento aumenta con el incremento de la tensión aplicada entre puerta y fuente, V_{GS}.

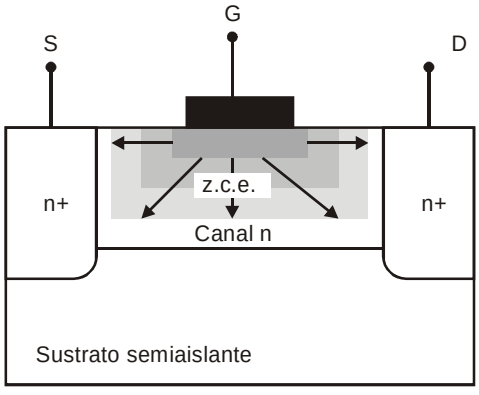


Figura 7.2.1 Estructura de un MESFET y detalle de la modificación de la zona de carga espacial para distintas tensiones aplicadas entre los contactos de puerta y fuente.

Puede ocurrir que al ir aumentando esta tensión se agote por completo el canal, anulándose la corriente entre drenador y fuente para tensiones superiores. A la tensión V_{GS} que agota el canal se le conoce con el nombre de tensión umbral V_t, al igual que se definió en el MOSFET. Para calcularla basta con igualar el espesor de la zona de carga espacial *h*, que viene dado por

$$h = \sqrt{\frac{2\varepsilon_s(\psi_0 - V_{GS})}{q N_D}}, \qquad (7.1)$$

con el espesor del canal *a* (Figura 7.2.2). Despejando de esa igualdad la tensión umbral se obtiene:

$$V_t \equiv V_{GS}(h=a) = \psi_0 - \frac{qa^2 N_D}{2\varepsilon_s}, \qquad (7.2)$$

donde N_D es la concentración de impurezas donadoras. Se suele definir también la tensión de agotamiento o "pinch-off", V_p, como:

$$V_p = \frac{qa^2 N_D}{2\varepsilon_s}. \qquad (7.3)$$

El espesor de la zona de carga espacial se puede rescribir:

$$h = a\sqrt{\frac{(\psi_0 - V_{GS})}{V_p}}. \qquad (7.4)$$

La unión metal semiconductor está polarizada en inverso para aislar el terminal de puerta de los otros terminales. Al modificar la zona de carga espacial con variaciones de la tensión V_{GS} aparecen efectos capacitivos entre la puerta y el canal, comportamiento análogo al resto de los transistores de efecto campo. Las diferencias con el resto de transistores de efecto campo estriban en (i) la localización del canal, (ii)

cómo se consigue el aislamiento de puerta y (iii) el tipo de material utilizado.

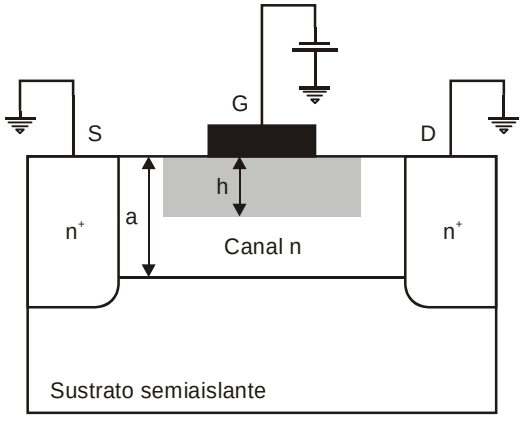


Figura 7.2.2 Definición del espesor del canal *h* en un MESFET.

7.3 Principio de operación

La conductancia que presentaría el canal tipo N sin tener en cuenta la zona de carga espacial sería:

$$G_0 = \frac{qN_D\mu_n Wa}{L}, \tag{7.5}$$

donde L es la longitud del canal y W su profundidad. La conductividad real de este canal incluyendo los efectos de la zona de carga espacial sería:

$$G_{ds} = \frac{qN_D\mu_n W(a-h)}{L}. \tag{7.6}$$

Cuando la zona de carga espacial ocupe todo el canal $(a = h)$ la conductividad se hará nula.

Cuando, además de aplicar una tensión a la puerta, la tensión V_{DS} es distinta de cero la diferencia de potencial entre puerta y un punto x del canal variará a lo largo del mismo. Se debe dividir el canal en elementos de longitud dx, cada uno de los cuales presentará un espesor $(a-h(x))$ y una conductividad diferentes entre sí (Figura 7.3.1). El espesor de la zona de carga espacial en un punto x es:

$$h(x) = \sqrt{\frac{2\,\varepsilon_s(\psi_0 - V_{GS} + V(x))}{q\,N_D}}, \tag{7.7}$$

donde $V(x)$ es el potencial del canal en ese punto referido al terminal de fuente. Este procedimiento se conoce como aproximación de canal

gradual, y es el que se ha venido usando para calcular la característica corriente tensión en todos los transistores de efecto campo. El primero que usó este método fue Shockley en su trabajo de 1952 "Un transistor de efecto campo unipolar" [1].

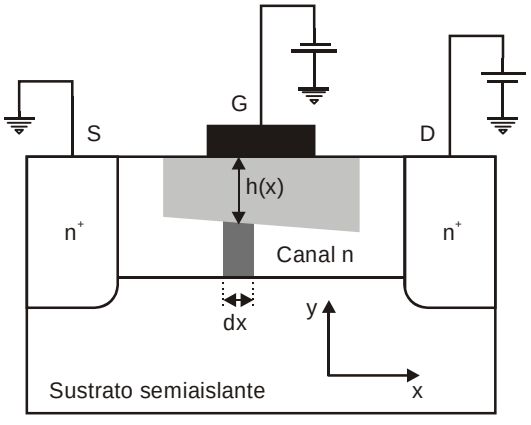


Figura 7.3.1 Variación del espesor del canal *a-h* en un MESFET por efecto de la tensión V_{DS} .

La resistencia de ese elemento de canal se puede escribir como:

$$dR = \frac{dx}{qN_D W\mu_n (a - h(x))}, \tag{7.8}$$

con lo que la caída de potencial en este pequeño segmento será:

$$dV = I_D \frac{dx}{qN_D W\mu_n (a - h(x))}. \tag{7.9}$$

Integrando esta ecuación entre 0 y L se obtiene la relación entre la corriente que circula por el canal con las tensiones aplicadas entre los terminales del transistor:

$$I_D = G_0 \left[V_{DS} - \frac{2}{3} \frac{(V_{DS} + \psi_0 - V_{GS})^{\frac{3}{2}} - (\psi_0 - V_{GS})^{\frac{3}{2}}}{(V_p)^{\frac{1}{2}}} \right]. \tag{7.10}$$

Esta aproximación es sólo válida mientras no aparezca el agotamiento del canal en la región de drenador, es decir, mientras no se cumpla $h(L) = a$. A partir de ese punto la corriente tomaría un valor constante que se obtiene imponiendo la condición de agotamiento $V_{DSsat} = V_p - \psi_0 + V_{GS}$:

$$I_D = G_0 \left[\frac{V_p}{3} + \frac{2}{3} \frac{(\psi_0 - V_{GS})^{\frac{3}{2}}}{(V_p)^{\frac{1}{2}}} - \psi_0 + V_{GS} \right]. \tag{7.11}$$

Si quisiéramos mantener una corriente finita en un canal de espesor nulo obligaría a trabajar con una velocidad de los electrones infinita o con un campo eléctrico infinito. Para solucionar este problema se puede hacer uso del hecho que la velocidad de deriva de los electrones se satura para campos elevados, como se observa en la Figura 7.3.2.

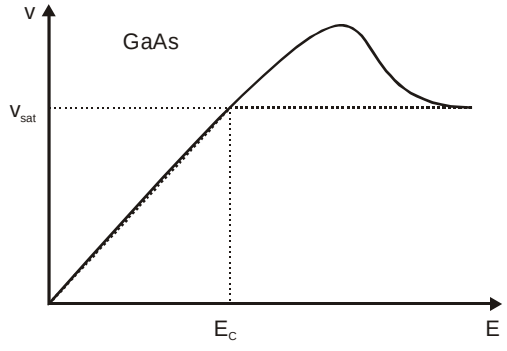


Figura 7.3.2 Relación entre la velocidad de los electrones en GaAs y el campo electrico aplicado al semiconductor.

Modelo de dos tramos de la velocidad

En los transistores de efecto campo de GaAs actuales es normal que se supere el campo eléctrico crítico, E_c, a partir del cual se satura la velocidad. En consecuencia, vamos a considerar que la expresión de la corriente (7.10) es válida para tensiones V_{DS} inferiores a V_{DSsat}, donde V_{DSsat} se define como la diferencia de potencial entre drenador y fuente a la cual el campo eléctrico en el drenador $E(L)$ se iguala al campo crítico E_c.

El campo eléctrico lateral se puede obtener de la siguiente relación:

$$\begin{aligned} J(x) &= \sigma E(x), \\ \frac{I_D}{W(a-h(x))} &= qN_D\mu_n E(x), \\ E(x) &= \frac{I_D}{qN_D\mu_n W(a-h(x))}. \end{aligned} \tag{7.12}$$

El campo es máximo en el extremo de drenador, $(x = L)$:

$$\begin{aligned} E(L) &= \frac{I_D(V_{DS})}{qN_D\mu_n W\left(a-\sqrt{\frac{2\varepsilon_s(\psi_0 - V_{GS}+V_{DS})}{qN_D}}\right)} \\ &= \frac{I_D(V_{DS})}{qN_D\mu_n Wa\left(1-\sqrt{\frac{\psi_0 - V_{GS}+V_{DS}}{V_p}}\right)}. \end{aligned} \tag{7.13}$$

Igualando el campo en el extremo de drenador al campo crítico $E(L) = E_c$ se tiene el inicio de la región de saturación $(V_{DS} = V_{DSsat})$. Evaluando la expresión de la corriente (7.10) para V_{DSsat} y haciendo uso de las definiciones:

$$u_{sat} \equiv \frac{V_{DSsat}}{V_p}, \quad u_g \equiv \frac{\psi_0 - V_{GS}}{V_p}, \quad \gamma \equiv \frac{E_c L}{V_p}, \tag{7.14}$$

se puede escribir la ecuación (7.13) de la forma:

$$\gamma = \frac{u_{sat} - \frac{2}{3}\left[(u_g + u_{sat})^{3/2} - u_g^{3/2}\right]}{1 - (u_g + u_{sat})^{1/2}}, \tag{7.15}$$

de donde se puede extraer el valor de V_{DSsat}.

Para $\gamma >> 1$ la solución de esta ecuación se puede aproximar por $u_{sat} + u_g = 1$, idéntico al modelo de canal gradual.

Para $\gamma << 1$ se puede aproximar $u_{sat} \approx \gamma$.

También se puede resolver numéricamente la ecuación (7.15) y obtener una fórmula de interpolación. Shur [2] demostró que dicha fórmula se podía aproximar por:

$$u_{sat} \approx \frac{\gamma(1 - u_g)}{\gamma + 1 - u_g}. \tag{7.16}$$

La corriente de saturación vendrá dada por:

$$\begin{aligned} I_D &= q v_{sat} W N_D (a - h(L)) \\ &= G_o L E_c \left(1 - \left(\frac{\psi_0 - V_{GS} + V_{DSsat}}{V_p}\right)^{1/2}\right) \\ &= G_o V_p \gamma (1 - (u_{sat} + u_g)^{1/2}). \end{aligned} \tag{7.17}$$

En el caso límite $\gamma \to \infty$ (dispositivo largo con pequeña tensión de estrangulamiento) la corriente de saturación se reduce a la obtenida con el modelo de canal gradual (7.11). En el otro extremo, $\gamma << 1$ (puerta corta o gran tensión de estrangulamiento), se puede aproximar la corriente de saturación (7.17) por:

$$I_D \approx G_o V_p \gamma (1 - (\gamma + u_g)^{1/2}). \tag{7.18}$$

Para valores intermedios Shur [3] propuso una formula de interpolación:

$$I_D = G_o V_p \gamma \frac{\left(1 - u_g\right)^2}{1 + 3\gamma}, \tag{7.19}$$

que también se puede escribir como:

$$I_D = \beta (V_{GS} - V_t)^2 , \tag{7.20}$$

donde

$$\beta = \frac{2\varepsilon_s \mu_n v_{sat} W}{a(\mu_n V_p + 3 v_{sat} L)}, \quad V_t = \psi_0 - V_p . \tag{7.21}$$

Ejemplo 7.1

Considerar un MESFET de GaAs con los siguientes parámetros: ψ_0 = 0.6 V, $N_D = 3\times10^{17}$ cm^{-3}, W = 20 μm, L = 1 μm, V_{GS} = 0 V. Considerar que la movilidad de los electrones en este material es 4000 cm^2/(Vs) y la velocidad de saturación 10^7 cm/s. Calcular la curva $I_{DS}-V_{DS}$ empleando el modelo de canal gradual y el modelo de dos tramos para la velocidad. Evaluar en este segundo caso la anchura del canal en la región de saturación.

Solución.

Con estos datos se puede calcular la tensión de agotamiento V_p, que toma el valor de 2.069 V, y el campo crítico 2.5×10^3 V/cm. Con el modelo de canal gradual la región de saturación comenzaría para

$$V_{DSsat} = V_{GS} - V_t = V_{GS} + V_p - \psi_0 = 1.469 \text{ V}$$

Si la relación velocidad-campo eléctrico la aproximamos por dos tramos lineales como se observa en la Figura 7.3.2 e igualamos el campo $E(L)$ de la expresión (7.13) a 2.5×10^3 V/cm se puede obtener el valor de la tensión V_{DS} que cumple estas condiciones: V_{DSsat} = 0.223 V. Si evaluamos $E_c \times L$ se obtiene 0.25V. Se puede comprobar que para $E_c \times L >> V_p$ se cumple $V_{DSsat} = V_t$ y en el caso $E_c \times L << V_p$ se cumple que $V_{DSsat} \approx E_c \times L$. En la Figura 7.3.3 se representa la curva I_D-V_{DS} para este transistor admitiendo (a) que la velocidad de los electrones crece de manera indefinida y (b) que la velocidad de los electrones se satura a partir del campo crítico (modelo de dos tramos para la velocidad).

Cuando comienza la saturación del transistor el canal se agota en el extremo de drenador. Sin embargo, para evitar hablar de espesores de canales nulos y de campos y velocidades infinitas lo más adecuado es hablar de un espesor del canal finito. Este espesor, b_{sat}, se puede estimar introduciendo la velocidad de saturación de los electrones:

$$I_{DSsat} = qN_D W b_{sat} v_{sat}$$

En este ejemplo I_{DSsat} =3.547 mA con lo que b_{sat} =0.037 μm. Se puede comparar este valor con el espesor del canal en la región de fuente: 0.046 μm (Figura 7.3.4).

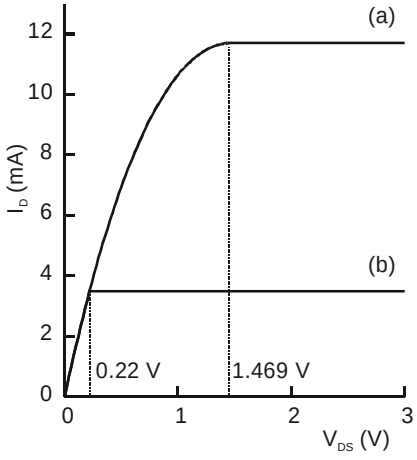


Figura 7.3.3 Representación de las curvas I_D–V_{DS} con el modelo de canal gradual y admitiendo la saturación de la velocidad de los electrones en el canal.

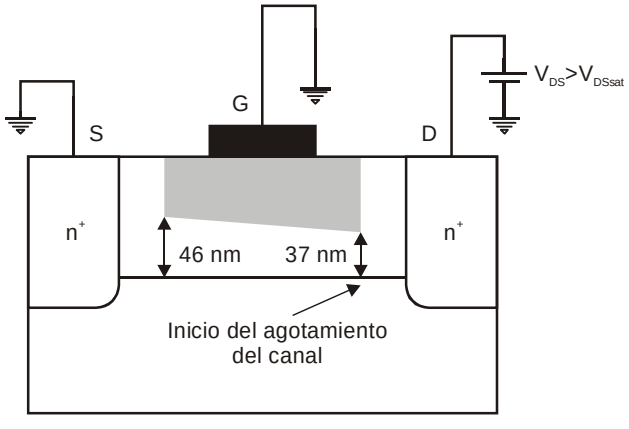


Figura 7.3.4 Estimación del espesor del canal en las regiones de fuente y drenador.

Modelo de Curtice

La conductancia del canal para valores bajos de la tensión V_{DS}, se puede obtener derivando la expresión de la corriente (7.10) con respecto a V_{DS} manteniendo constante V_{GS}:

$$g_{chi} = \left(\frac{\partial I_D}{\partial V_{DS}} \right)_{V_{GS}=cte} = G_0 \left[1 - \left(\frac{\psi_0 - V_{GS} + V_{DS}}{V_p} \right)^{1/2} \right]. \qquad (7.22)$$

El modelo del dispositivo que se emplea en diseño asistido por ordenador debe reproducir las curvas I-V en todo el rango de tensiones, no solo en saturación. Curtice [4] propuso una expresión que interpolaba la corriente en todo el rango de tensiones:

$$I_D = I_{DSs}(1+\lambda V_{DS})\tanh(\eta V_{DS}), \tag{7.23}$$

donde

$$I_{DSs} = \beta(V_{GS} - V_t)^2, \tag{7.24}$$

$$I_{DSs} = qWv_{sat}N_D a(1-\frac{\psi_0 - V_{GS}}{V_p}), \tag{7.25}$$

η es un parámetro que se elige de manera que cuando $V_{DS} \to 0$ el valor de la corriente de esta última ecuación converja al valor de la corriente del modelo de canal gradual:

$$\eta = \frac{g_{chi}}{I_{DSs}} \tag{7.26}$$

y λ es un parámetro empírico que da idea de la conductancia de salida y refleja la modulación de la longitud del canal.

Efectos parásitos

Las resistencias serie de fuente y drenador (Figura 7.3.5) juegan un papel importante en las características I-V del transistor. La resistencia serie de fuente representa la resistencia total del contacto óhmico de fuente y la resistencia de la zona neutra del semiconductor entre el contacto de fuente y la parte activa del canal. La resistencia serie de drenador refleja o mismo que la de fuente pero referida a la región de drenador.

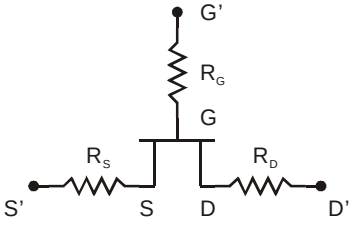


Figura 7.3.5 Resistencias serie en los tres terminales del transistor.

Las diferencias de potencial entre los terminales externos se pueden relacionar con las del dispositivo intrínseco:

$$\begin{aligned} V_{G'S'} &= V_{GS} + I_D R_S, \\ V_{D'S'} &= V_{DS} + I_D(R_S + R_D). \end{aligned} \tag{7.27}$$

Para valores $V_{DS} << V_{DSsat}$ podemos escribir la corriente de drenador

$$I_D \approx g_{chi} V_{DS} \approx g_{ch} V_{D'S'}, \tag{7.28}$$

donde g_{ch} es la conductancia del canal extrínseca. Combinando las ecuaciones (7.27) y (7.28) se encuentra la relación:

$$g_{ch} = \frac{g_{chi}}{1 + g_{ch}(R_S + R_D)}. \tag{7.29}$$

Despejando el valor de V_{GS} de (7.27) e introduciéndolo en (7.24) se obtiene una nueva expresión para la corriente de saturación, I_{DSs}, en función de las tensiones de los terminales externos:

$$I_{DSs} = \frac{1 + 2\beta R_S (V_{G'S'} - V_t) - (1 + 4\beta R_S (V_{G'S'} - V_t))^{1/2}}{2\beta {R_S}^2}. \tag{7.30}$$

Teniendo en cuenta estos elementos extrínsecos se puede rescribir la relación corriente tensión (7.23):

$$I_D = I_{DSs}(1 + \lambda V_{D'S'}) \tanh(\eta V_{D'S'}), \tag{7.31}$$

$$\eta = \frac{g_{ch}}{I_{DSs}}. \tag{7.32}$$

Hasta ahora se ha considerado que la corriente que circula por la puerta es despreciable pues la unión Schottky está polarizada en inversa. Sin embargo, hay situaciones en las que puede ser útil hacer circular corriente por la puerta, en particular cuando se está interesado en extraer parámetros del dispositivo. Para tener en cuenta estas situaciones se suele modelar la corriente de puerta mediante la combinación de dos diodos. Uno de ellos está conectado entre el contacto de puerta y el de fuente y el otro entre el contacto de puerta y drenador (Figura 7.3.6). Ello daría lugar a una corriente de puerta de valor:

$$\begin{aligned} I_G &= I_{gs} + I_{gd}, \\ I_{gs} &= I_{g0} \times \exp\left(q \frac{V_{G'S'} - I_G R_G - (I_D + I_{gs}) R_S}{\eta KT} \right), \\ I_{gd} &= I_{g0} \times \exp\left(q \frac{V_{G'D'} - I_G R_G - (I_D + I_{gd}) R_D}{\eta KT} \right). \end{aligned} \tag{7.33}$$

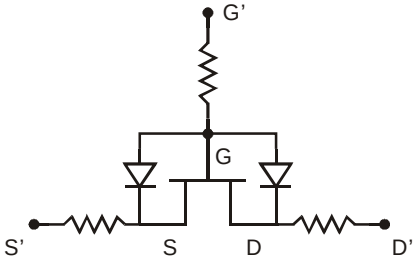


Figura 7.3.6 Inclusión de las uniones Schottky en el modelo del transistor.

Extracción de parámetros del MESFET

Las ecuaciones estudiadas en el modelo de Curtice junto a las correcciones por los elementos parásitos forman un conjunto completo de expresiones que se emplean en el modelo analítico del MESFET usado en el simulador de circuitos de GaAs UM-SPICE. Los parámetros del modelo están relacionados con la geometría, el dopado y los parámetros del material tales como la velocidad de saturación y la movilidad de bajo campo.

Los parámetros de un MESFET real se pueden extraer a partir de medidas experimentales corriente-tensión.

- El parámetro λ se puede extraer de una curva $I_D - V_{DS}$ para un valor constante de $V_{G'S'}$.
- Las corrientes de saturación se extrapolan a V_{DS} =0 para todos los valores de $V_{G'S'}$, considerando conocido el valor de λ.
- Para obtener R_S, se representa $I_D^{1/2}$ en saturación en función de ($V_{G'S'} - I_{DS}R_S$) para diferentes valores de R_S. Esta representación será una línea recta si se ha elegido el valor adecuado para R_S. Se puede tomar como criterio el que el que el coeficiente de correlación sea mayor que un valor determinado.
- De la pendiente de esta recta y del corte con el eje de abcisas se obtienen β y V_T respectivamente.
- El potencial barrera ψ_0 se determina a partir de las características I-V de la puerta.
- El espesor del canal y el dopado se pueden extraer combinando (7.3) y la dosis de implantación (producto $N_D \times a$).
- Considerando que las resistencias R_D y R_S son iguales se puede obtener la resistencia intrínseca del canal R_i a partir de la pendiente de la curva $I_D - V_{DS}$ en la región lineal para V_{GS} elevada:

$$R_{chi} = R_{DS} - R_S - R_D , \tag{7.34}$$

donde R_i se relaciona con la tensión de puerta $R_{chi} = 1/g_{chi}$:

$$R_{chi} = \frac{L}{q\mu_n N_D W(a-h)} = \frac{L}{qa\mu_n N_D W(1-((\psi_0 - V_{GS})/V_p)^{1/2})} . \tag{7.35}$$

- De esta ecuación se puede obtener la movilidad de los electrones μ_n.
- Posteriormente se puede calcular la velocidad de saturación despejándola de la definición de β.

7.4 Circuito equivalente de pequeña señal

Los transistores de efecto campo, en particular los MESFETs de GaAs, son útiles como amplificadores de bajo ruido, como generadores de potencia con un alto rendimiento, y en aplicaciones lógicas de alta velocidad. El modelo equivalente del transistor se muestra en la Figura 7.4.1, donde se sitúa cada elemento en la estructura real.

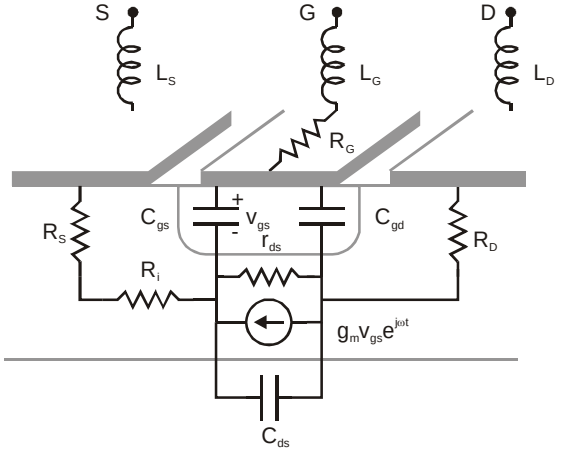


Figura 7.4.1 Localización de los elementos del modelo de pequeña señal en el MESFET.

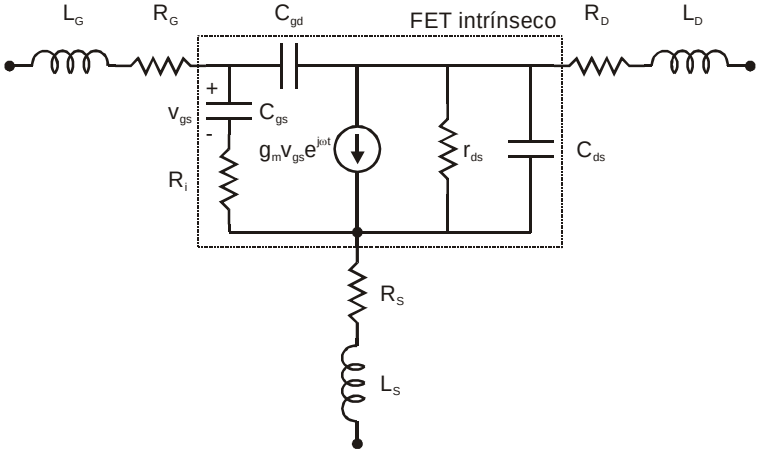


Figura 7.4.2 Modelo de pequeña señal del MESFET.

El MESFET (Figura 7.4.2) se puede dividir en dos partes: el dispositivo intrínseco y los elementos extrínsecos o parásitos. La parte intrínseca representa la zona activa del dispositivo, aquella cuyas características dependen de la tensión aplicada a los terminales. El funcionamiento del dispositivo intrínseco es el que se acaba de describir

en el apartado anterior e incluye la zona activa del canal. Los elementos extrínsecos no son necesarios para que el dispositivo pueda operar con normalidad y no varían con las condiciones de polarización. Sin embargo, en aplicaciones de alta frecuencia y alta velocidad el efecto de las partes reactivas del MESFET puede deteriorar el funcionamiento del mismo por lo que hay que incluirlos en el modelo equivalente.

El modelo está basado en la estructura real del dispositivo y es válido hasta la región de microondas (varias decenas de gigahercios). Cada elemento refleja las peculiaridades de alguna región del dispositivo. C_{gs} y C_{gd} representan el almacenamiento de carga en las zonas de carga espacial de las regiones puerta-fuente y puerta-drenador. La resistencia R_S representa la resistencia total del contacto óhmico de fuente y la resistencia de la zona neutra del semiconductor entre el contacto de fuente y la parte activa del canal. La resistencia R_D igual que la de fuente pero referida a este terminal. L_G, L_S y L_D representan las inductancias de puerta, fuente y drenador respectivamente. R_G representa la resistencia de la metalización de puerta. R_i es la resistencia del canal entre puerta y fuente.

Los valores de los elementos extrínsecos e intrínsecos dependen de la estructura del canal, de la concentración de impurezas, del tamaño del dispositivo, de su "layout" y de los procesos de fabricación. Los principales parámetros intrínsecos son la transconductancia g_m, la capacidad de entrada C_{gs}, la resistencia de salida r_{ds} y la capacidad de realimentación C_{gd}. Estos parámetros dependen de las tensiones de polarización del dispositivo. El retardo τ de la corriente de drenador o de la transconductancia respecto a la señal de entrada refleja el tiempo que necesitan los electrones en atravesar la zona activa del canal. Es el tiempo empleado para el intercambio de carga con la zona de vaciamiento en la región de saturación del canal. Es cero hasta que se alcanza la velocidad de saturación. Se espera que aumente con la tensión de drenador y disminuya cuando aumente la tensión de puerta.

Estimación de los elementos del circuito

- Transconductancia: Derivando la expresión de la corriente de drenador en la zona lineal (7.10) (obtenida del modelo de canal gradual) se obtiene para este parámetro:

$$g_m = \left(\frac{\partial I_D}{\partial V_{GS}} \right)_{V_{DS}=cte} = G_0 \left[\frac{(V_{DS} + \psi_0 - V_{GS})^{\frac{1}{2}} - (\psi_0 - V_{GS})^{\frac{1}{2}}}{(V_p)^{\frac{1}{2}}} \right]. \qquad (7.36)$$

Para valores bajos de la tensión de drenador $(V_{DS} << \psi_0 - V_{GS})$ la corriente de drenador y la transconductancia se pueden aproximar por las siguientes expresiones:

$$I_D \approx G_0\left[1-\left(\frac{\psi_0 - V_{GS}}{V_p}\right)^{1/2}\right]V_{DS},$$
$$g_m \approx \frac{G_0 V_{DS}}{2(V_p)^{1/2}(\psi_0 - V_{GS})^{1/2}}. \qquad (7.37)$$

Derivando la expresión de la corriente en saturación (7.11) se obtiene la transconductancia en esta región:

$$(g_m)_{sat} \approx G_0\left[1-\left(\frac{\psi_0 - V_{GS}}{V_p}\right)^{1/2}\right]. \qquad (7.38)$$

- Capacidades: Las cargas almacenadas en la estructura dan lugar a efectos capacitivos. La zona de carga espacial de la región de puerta varía en espesor a lo largo del canal por lo que el acoplamiento capacitivo entre el metal de puerta y el semiconductor se encuentra distribuido. En la práctica esta capacidad se puede representar mediante dos capacidades asociadas a las uniones Schottky de drenador-puerta y fuente-puerta.

$$C_{GS} = \frac{C_{g0}}{\sqrt{1-\frac{V_{GS}}{\psi_o}}},$$
$$C_{GD} = \frac{C_{g0}}{\sqrt{1-\frac{V_{GD}}{\psi_o}}}, \qquad (7.39)$$

donde

$$C_{g0} = \frac{WL}{2}\sqrt{\frac{qN_D\varepsilon_s}{2\psi_0}}. \qquad (7.40)$$

- Resistencia intrínseca del canal entre puerta y fuente, $r_{ds} = R_{chi} = 1/g_{chi}$.

$$r_{ds} = \frac{L}{q\mu_n N_D W(a-h)} = \frac{L}{qa\mu_n N_D W\left\{1-\left[(\psi_0 - V_{GS})/V_p\right]^{1/2}\right\}}. \qquad (7.41)$$

- Resistencia asociada a la región neutra entre los contactos de fuente y puerta R_i (Figura 7.4.3). Admitiendo que la longitud de esta región es igual a la mitad de la longitud del canal podemos escribir:

$$R_i = \frac{\rho_s (L/2)}{Wa} = \frac{(L/2)}{q\mu_n N_D Wa}. \qquad (7.42)$$

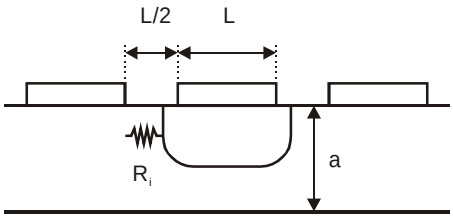


Figura 7.4.3 Resistencia asociada a la región neutra entre los contactos de fuente y puerta.

- Capacidad drenador fuente C_{ds} (Figura 7.4.4): refleja el acoplo capacitivo entre los contactos de fuente y drenador a través del sustrato semiaislante

$$C_{ds} = \frac{\varepsilon_s A}{X} = \frac{\varepsilon_s aW}{3L}. \qquad (7.43)$$

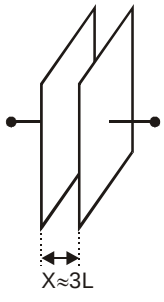


Figura 7.4.4 Capacidad asociada al acoplo entre los contactos de drenador y fuente a través del sustrato.

Un circuito simplificado que permite predecir de forma muy acertada el funcionamiento de los transistores de efecto campo de GaAs en circuitos de microondas como amplificadores y osciladores se muestra en la **Figura 7.4.5**. En él no aparecen las resistencias de los contactos de fuente, drenador y puerta así como la capacidad drenador puerta.

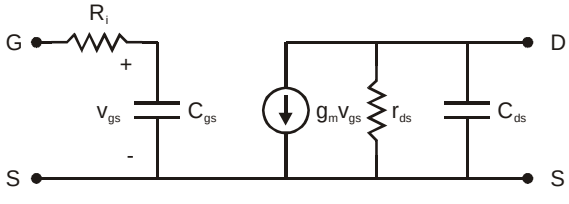


Figura 7.4.5 Modelo de pequeña señal simplificado para el MESFET.

Los parámetros de un MESFET típico suelen tomar los siguientes valores: C_{gs} =0.3 pF, C_{ds} =0.05 pF, g_m =40 mS, r_{ds} =600 Ω, R_i =25 Ω.

Como aplicación de este circuito equivalente se puede calcular la ganancia disponible máxima del MESFET, que se define como la ganancia de potencia máxima que podemos conseguir a cualquier frecuencia admitiendo que la entrada y salidas del circuito están adaptadas. Para ello se considera el circuito de la Figura 7.4.6.

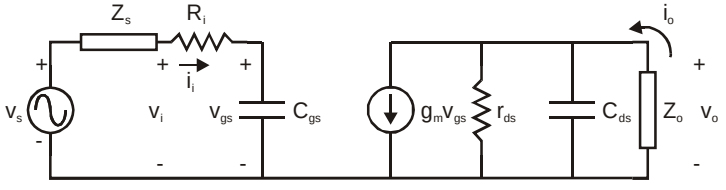


Figura 7.4.6 Modelo de pequeña señal del MESFET excitado con una fuente de tensión a la entrada y cargado en su puerta de salida con una carga Z_o.

En primer lugar se calcula el valor de las impedancias Z_s y Z_o que permiten adaptar la entrada y la salida:

$$Z_s = R_i + \frac{j}{C_{gs}\omega}, \qquad Z_O = \frac{1}{1/r_{ds} - j\omega C_{ds}}. \tag{7.44}$$

Una vez conseguido esto se calculan las potencias entregadas a la entrada del dispositivo y a la carga Z_o:

$$P_i = 1/2 \times \mathrm{Re}(I_i^* V_i), \qquad P_o = 1/2 \times \mathrm{Re}(I_o^* V_o). \tag{7.45}$$

Su cociente proporciona la ganancia de potencia máxima:

$$MAG = \frac{1}{4}\frac{g_m^2}{(2\pi)^2}\frac{r_{ds}}{R_i}\frac{1}{(C_{gs} f)^2}. \tag{7.46}$$

Introduciendo en esta ecuación las expresiones de los distintos parámetros que en ella aparecen (7.37)-(7.42) se tiene que la ganancia máxima es proporcional a N_D^2 e inversamente proporcional al producto $(L \times f)^2$. Esto indica que si queremos aumentar el rango de operación de este dispositivo debemos disminuir la longitud del canal o aumentar el dopado del semiconductor. Las longitudes de canal que se emplean hoy en día son inferiores a las 0.5 μm, valor que ronda el límite inferior para la fotolitografía óptica. Con otras técnicas como la litografía por haces de electrones se pueden conseguir longitudes inferiores a las 0.2 μm. Con

respecto al aumento de las impurezas del canal semiconductor éstas incrementan los mecanismos de dispersión por impurezas ionizadas, disminuyendo la movilidad, por lo que no es una solución acertada el aumento indiscriminado de este parámetro.

RESUMEN

En este capítulo se ha introducido al transistor de efecto campo metal semiconductor (MESFET). Se ha hablado de la importancia del material sobre el que normalmente se construye este dispositivo (GaAs) y de las razones por las que se utiliza. Se ha descrito la estructura y el funcionamiento básico. Se han estudiado diferentes modelos que tienen como objeto el llegar a una expresión para la corriente de drenador en función de las tensiones aplicadas a los terminales. Se ha añadido complejidad a estas expresiones al introducir otros efectos como las resistencias parásitas o la corriente de puerta. Se ha propuesto un método para extraer los parámetros característicos del dispositivo. Finalmente se ha presentado el modelo de pequeña señal y se ha estimado el valor de los parámetros que en él aparecen.

CUESTIONES Y PROBLEMAS

1. Considerando el circuito con un MESFET de la Figura 7.4.6 calcular la ganancia de potencia disponible máxima y la ganancia en potencia en caso de que la impedancia de fuente y carga tomen el valor $Z_S = Z_o = 50\ \Omega$. Calcular el cuadrado de la ganancia en corriente en cortocircuito. En este tercer caso cambiar la fuente de tensión por una de corriente y considerar $Z_o = 0$.

   Para los parámetros del modelo de pequeña señal utilizar los del transistor MA4TF5005 polarizado a V_{DS}=3V e I_D=10 mA (R_i=5 Ω, r_{ds}=406 Ω , C_{gs} = 0.35 pF, C_{ds} = 0.16 pF, g_m = 32 mS). Evaluar dichas ganancias comparándolas entre sí en una gráfica.

2. Considérese el dispositivo cuya estructura se representa en la Figura P.1, con los siguientes valores de los parámetros:

   $a = 2\ \mu m$, $L = 20\ \mu m$, $W = 10\ \mu m$, $N_D = 10^{16}\ cm^{-3}$

   Para este dispositivo, conocido como MESFET, son aplicables las expresiones obtenidas en el análisis de JFET con unión abrupta. Las difusiones N^+ realizadas debajo de la fuente y el drenador se realizan exclusivamente para conseguir contactos óhmicos, mientras que el contacto metálico de puerta es rectificador. Si la barrera de este contacto vista desde el metal es de 0.6 eV, y

despreciamos el abatimiento Schottky de la barrera debido al efecto imagen, calcular:

a) El potencial barrera visto desde el semiconductor.

b) La tensión de puerta que agota el canal. (Para poder aplicar la misma expresión que se utiliza en el JFET admitir que el metal es como si fuera un semiconductor fuertemente dopado).

c) La transconductancia del canal.

d) El tiempo de tránsito a través del canal para una tensión de drenador tal que se alcance la saturación con V_G=0.5 V_p.

e) La limitación en frecuencia del dispositivo.

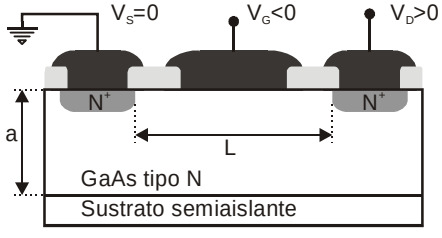


Figura P.1.

REFERENCIAS

[1] W. Shockley. "A unipolar field effect transistor", *PROC. IRE*, vol. 40, pp. 1365-1376, November 1952.

[2] M.S. Shur. “Low field mobility, effective saturation velocity and performance of submicron GaAs MESFETs ” *Electron., Lett.*, Vol. 18 (21), pp. 909-911, Oct. 1982.

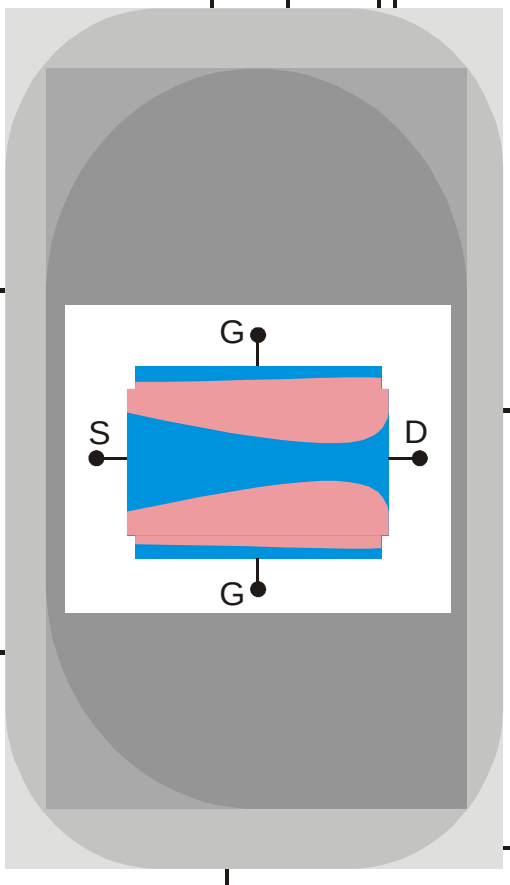
[3] M.S. Shur. “Analitical model of GaAs MESFETs”, *IEEE Trans. Electron Devices*, vol. ED-25, pp. 612-618, June 1978.

[4] W.R. Curtice. “A MESFET model for use in the design of GaAs integrated circuits”, *IEEE Trans. MTT*, vol. MTT-29, pp. 448-456, May 1980.

Capítulo 8

Dispositivos optoelectrónicos.

Diodo Láser



Índice

Objetivos

- Definir y clasificar funcionalmente los dispositivos optoelectrónicos.
- Describir las características de la interacción luz-materia, las transiciones entre estados y la Ecuación de Einstein.
- Definir y evaluar los coeficientes de ganancia y absorción de la intensidad de la radiación luminosa.
- Establecer las condiciones de conservación de la energía y el vector de onda en las transiciones entre estados en materiales semiconductores.
- Describir los materiales semiconductores utilizados en la fabricación de estos dispositivos.
- Determinar las condiciones umbral en los diodos láser de homounión.

Palabras Clave

Dispositivos optoelectrónicos.
Transiciones entre estados. Ecuación de Einstein
Absorción de fotones.
Emisión espontánea y estimulada de fotones.
Coeficiente de absorción y de ganancia de la intensidad de radiación
Conservación de la energía y del momento cristalino.
Transiciones entre bandas de energía.
Amplificadores ópticos.
Láseres semiconductores.
Condición de mantenimiento de la oscilación.
Condición umbral de oscilación.
Densidad conjunta de estados.
Perfil de emisión espontánea en un semiconductor.
Espectro de emisión estimulada
Diodos láser de homounión

8.1 Introducción.

Dispositivos optoelectrónicos semiconductores
Son aquellos dispositivos semiconductores cuyo objetivo es obtener radiación luminosa a partir de corriente eléctrica o bien generar corrientes eléctricas a partir de radiación luminosa. Aunque hay dispositivos optoelectrónicos que no son semiconductores, para nosotros dispositivos optoelectrónicos equivaldrá a dispositivos optoelectrónicos semiconductores

Hasta ahora todos los dispositivos estudiados tienen como mecanismo básico de su funcionamiento el transporte de corriente por uno o dos tipos de portadores de carga, electrones o huecos (o ambos). Existe otros tipos de dispositivos semiconductores cuya función es, en unos casos, producir luz a partir de corrientes eléctricas, en otros, generar corrientes eléctricas a partir de la incidencia de luz sobre ellos: son los denominados Dispositivos Optoelectrónicos Semiconductores. Aunque existen dispositivos realizados con materiales no semiconductores que pueden realizar esta misma función (por ejemplo, los fotomultiplicadores) para nosotros, y en el contexto de esta materia, cuando hablemos de dispositivos optoelectrónicos nos referiremos siempre a dispositivos optoelectrónicos realizados con semiconductores.

Por lo tanto, tenemos que introducir un nuevo elemento a lo ya conocido: la radiación luminosa, y estudiaremos algunos efectos de la interacción de la luz con la materia, y de forma especial con los semiconductores y su acción en los dispositivos con ellos fabricados. La importancia actual de este tipo de dispositivos es obvia para cualquiera que esté un poco familiarizado con los circuitos electrónicos, e incluso con la electrónica doméstica: los mandos a distancia de los electrodomésticos habituales funcionan (en su gran mayoría) por emisión y recepción de rayos infrarrojos. El origen de la información que se transmite por fibras ópticas, propias de redes de comunicación avanzadas, suele estar en la emisión de luz por un dispositivo LED o láser semiconductor. Los optoacopladores, en circuitos de control e instrumentación y las células solares son otros componentes en los que la interacción de la luz con las propiedades de uniones p-n determina el funcionamiento de estos dispositivos.

Además del aspecto aplicado de los dispositivos optoelectrónicos que acabamos de citar, también conviene señalar que desde un punto de vista de conocimiento básico de las propiedades de los materiales semiconductores, las experiencias realizadas iluminando de forma adecuada y controlada los semiconductores han ayudado en gran manera a la comprensión de las propiedades de los mismos. Así pues, tanto desde un punto de vista fundamental como aplicado, estos elementos revisten una importancia creciente en el mundo de la electrónica y merece la pena que nos detengamos en su estudio.

Vamos a hacer ahora una puntualización en cuanto a la nomenclatura: Por lo general, por "luz" entendemos la parte del espectro de la radiación electromagnética a la que es sensible el ojo humano, esto es, que "vemos". Con más propiedad, a esta zona deberíamos llamarla "espectro visible". Aquí, y en un abuso de lenguaje, por "luz" entenderemos radiación electromagnética en general, aunque los valores

numéricos los centraremos en el espectro visible y la zona próxima del infrarrojo -aquella que está limitando con el color rojo-. Alguna vez, aunque muy raramente, incluiremos el ultravioleta.

8.2 Clasificación de los dispositivos optoelectrónicos

Normalmente, los dispositivos optoelectrónicos se clasifican en función de que sean capaces de emitir luz por ellos mismos o respondan de determinada manera al recibirla. Los primeros se denominan *emisores de luz*, y dentro de ellos, tenemos los dispositivos LED y los LÁSER.

Los segundos, *receptores de luz*, son los que producen corriente eléctrica al recibir una radiación luminosa. En este grupo podemos distinguir dos tipos distintos, que tienen, por diseño y objetivos, características diferentes: unos cuya tarea primordial es detectar pequeñas intensidades de luz (la emisión de un mando a distancia, por ejemplo, o en casos más especializados y particulares, la débil luz de una estrella a través de un telescopio), y están formados por los dispositivos denominados fotodiodos y fototransistores, y otros que se dedican a producir tanta potencia eléctrica como sea posible: las células solares. Aún dentro de estos grupos se pueden establecer clasificaciones mas finas en función de su construcción o de variaciones en su forma de operar, que pueden ser importantes. Más adelante, al explicar los distintos tipos de estos dispositivos, veremos estas diferencias. Por ahora, para fijar ideas, basta con tener en cuenta el cuadro siguiente:

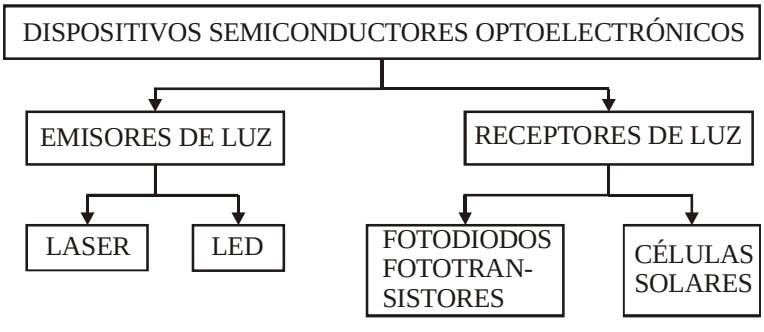


Figura 8.2.1 Clasificación de los dispositivos semiconductores optoelectrónicos según su función

8.3 Interacción luz-materia

Puesto que los dispositivos optoelectrónicos se basan en la interacción luz-materia, necesitamos introducir algunos conceptos básicos de estos procesos, manteniéndonos al nivel que necesitaremos. Desde luego, esta presentación será muy fenomenológica, sin entrar en sus fundamentos básicos, lo que nos llevaría a profundizar en la mecánica cuántica mucho más allá de lo razonable en este curso. Con todo, unas

nociones mínimas de esa materia serán necesarias, ya que no podremos evitarla por completo por la misma naturaleza de lo que aquí estamos tratando. Y para entrar en el tema, vamos ahora, en primer lugar, a recordar algunas definiciones y propiedades de las ondas electromagnéticas

Algunas propiedades de las ondas electromagnéticas.

Las características más elementales de una onda que se propaga en un medio material vienen dadas por su frecuencia, ν, y su longitud de onda λ. Estas magnitudes están relacionadas por

$$\lambda \cdot \nu = v \quad \text{con} \quad v = c/n\,, \tag{8.1}$$

donde v es la velocidad de propagación de la onda en el medio, lo que normalmente llamamos la velocidad de la luz en ese medio, *n* su índice de refracción y *c* la velocidad de la luz en el vacío (aproximadamente $3 \cdot 10^{10}$cm/s). Normalmente, la velocidad de propagación en el aire se toma igual a la del vacío, y así lo haremos nosotros (lo que equivale a tomar $n = 1$ en el aire). En general, *n* depende de la frecuencia, y se define un nuevo índice de refracción n_g que toma en cuenta dichas variaciones por

$$n_g = n + \nu \frac{dn}{d\nu}\,. \tag{8.2}$$

Este valor interviene en las fórmulas que se deducen desde principios básicos y que nosotros utilizaremos más adelante.

Conviene tener en cuenta que la frecuencia no cambia al pasar de un medio material a otro. Como la velocidad de propagación sí cambia, la longitud de onda también. Tendremos pues que tener en cuenta esta circunstancia y tener siempre presente a que medio nos referimos, que en general será o bien el semiconductor o bien el aire (que tomaremos como vacío).

El número de onda $\bar{\nu}$ se define como

$$\bar{\nu} = \frac{\nu}{c}; \qquad \text{en el vacío,} \qquad \bar{\nu} = \frac{1}{\lambda} \tag{8.3}$$

El vector de onda **k** es un vector que indica la dirección de propagación de la onda en el medio, y cuyo módulo se indica a continuación:

$$k = \frac{2\pi}{\lambda}; \qquad \text{en el vacío,} \quad k = 2\pi\bar{\nu} \tag{8.4}$$

De acuerdo con la física cuántica, la energía de una onda electromagnética se transmite en "paquetes" de valor $h\nu$, siendo *h* la constante de Plank. Cada uno de estos paquetes recibe el nombre de **fotones**, y la energía de una onda depende del "número de fotones que transporta la onda". En realidad, esta frase no tiene sentido, ya que lo que se puede medir es el número de fotones que *inciden* en una superficie en

un *tiempo dado*. Esto nos daría la energía que ha alcanzado la superficie en ese tiempo, que lógicamente, depende del tamaño de la superficie y del tiempo durante el que hemos medido. Para conseguir un parámetro que sea independiente del área y del tiempo de medida definimos la **intensidad** de la radiación como *la energía de la onda incidente por unidad de área y unidad de tiempo:*

$$I = \frac{Energía}{Area \cdot Tiempo} = \frac{Potencia}{Area}. \qquad (8.5)$$

A título de información, la potencia media aportada por el sol sobre la superficie terrestre, y en latitudes como las nuestras en el momento de máxima iluminación es de unos 1200 W/m^2 = 120 mW/cm^2 = 1.2 mW/mm^2.

Si consideramos un elemento de volumen δV en el espacio, atravesado por radiación luminosa, dentro de él hay almacenada una cierta cantidad de energía, procedente del campo electromagnético que forma la radiación. Si la radiación tiene un espectro continuo, llamemos $\rho(\nu)\cdot d\nu$ a la densidad de energía en el volumen (energía por unidad de volumen) en el intervalo de frecuencias entre ν y $\nu+d\nu$, e $I(\nu)\cdot d\nu$ a la intensidad de la radiación en el mismo intervalo. Se cumple que:

$$I(\nu)\cdot d\nu = \left[\rho(\nu)\cdot d\nu\right]\cdot \mathrm{v}\ , \qquad (8.6)$$

siendo v es la velocidad de propagación de la onda electromagnética en el medio en que esté el volumen considerado. En el caso de una onda monocromática, esto es, que tenga una sola frecuencia, I representa la intensidad de esa onda, y (8.6) queda:

$$I = \rho \cdot \mathrm{v}\ . \qquad (8.7)$$

Finalmente, nos quedan dos propiedades más de las ondas: su fase y su polarización. La fase tiene el sentido habitual, y la polarización indica las direcciones en las que vibran los campos eléctrico y magnético, y son propios de cada onda.

8.4 Transiciones entre estados. Ecuación de Einstein

A principios del siglo XX se hicieron las primeras hipótesis cuánticas con el fin de explicar la estructura y estabilidad de los átomos. Uno de los primeros modelos de éxito fue el de Bohr, que establecía que las transiciones de electrones entre dos niveles de energía E_2 y E_1 ($E_2 > E_1$) sólo se podían realizar si estaban acompañadas por la absorción o emisión de un fotón de energía $h\nu = E_2 - E_1$. En 1916, y con el fin de deducir de forma estadística la densidad de energía emitida por el cuerpo negro, A. Einstein propuso un modelo dinámico de las transiciones electrónicas que podían tener lugar entre dos estados, y, con respecto a la forma en que se emitían o absorbían los fotones, consideró los siguientes tipos:

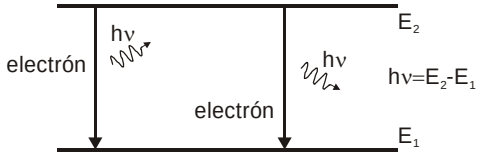


Figura 8.4.1 En la emisión espontánea los electrones pasan del nivel 2 al 1 de forma aleatoria, y los fotones que se emiten salen en todas direcciones y con cualquier fase en cualquier instante de tiempo

a) *Emisión espontánea de un fotón del nivel 2 al nivel 1.*

En este caso un electrón pasa del nivel E_2 al nivel E_1, y se emite un fotón con energía $h\nu = E_2 - E_1$ en cualquier dirección, con cualquier polarización y con fase aleatoria, sin relación con la de ningún otro fotón. Además, si N_2 designa la densidad (número por unidad de volumen) de átomos cuyo nivel 2 está ocupado por electrones, propuso que la variación temporal de la misma fuera de la forma:

$$\left.\frac{dN_2}{dt}\right)_{esp} = -A_{21} \cdot N_2 \,. \tag{8.8}$$

siendo A_{21} una constante que depende de los niveles particulares que estemos considerando. (Obsérvese que esta expresión presupone que cualquier electrón que pase del nivel 2 al 1 encuentra un estado vacío en el nivel 1, es decir, no tiene en cuenta el principio de exclusión de Pauli. Téngase en cuenta que en 1916 no se conocía dicho principio).

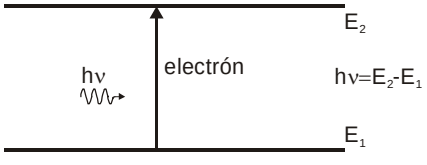


Figura 8.4.2 El fotón absorbido produce la transición del electrón del nivel 1 al 2. La desaparición del fotón produce una disminución de la intensidad de la radiación presente

b) Absorción de un fotón

Ahora el electrón pasa del nivel 1 al 2 mediante la absorción de un fotón de igual energía que el caso anterior. Para ello es necesario que haya fotones presentes, es decir, radiación procedente de alguna otra fuente (otros átomos, las paredes del recipiente, radiación externa al sistema, etc). Si $\rho(\nu)\cdot d\nu$ es la densidad de energía (por unidad de volumen) entre las frecuencias ν y $\nu + d\nu$, la variación temporal de la densidad de átomos N_1 cuyo nivel 1 está ocupado por electrones es de la forma

$$\left.\frac{dN_1}{dt}\right)_{abs} = -B_{12}\cdot\rho(\nu)\cdot N_1 = -\left.\frac{dN_2}{dt}\right)_{abs} \quad . \tag{8.9}$$

Como antes, B_{12} es otra constante de proporcionalidad.

c) *Emisión estimulada*

Los dos tipos *a)* y *b)* citados anteriormente podían corresponder más o menos con las ideas admitidas que ya formaban parte del modelo atómico de Bohr, y ser lógicas para la época. Pero Einstein propuso además otro tipo de proceso, las transiciones del nivel 2 al nivel 1 por emisión estimulada, en la que un fotón *presente* provocaría (*estimularía*) la transición generando otro fotón. Esta idea era completamente nueva para la época. La explicación que da la mecánica cuántica actual es bastante compleja, y hay que tener en cuenta que los fotones son partículas de spin entero denominadas bosones. Este tipo de partículas tienen la propiedad de que pueden haber muchas en un mismo estado (no rige para ellas el principio de exclusión de Pauli), y que cuantas más hay en un estado más probable es que se les añada otra más. Pero además, los fotones emitidos salen en el mismo estado que el que provoca la emisión, esto es, *salen con la misma energía, fase, dirección y polarización que el fotón que induce la transición*. Así, si un grupo de *n* fotones idénticos induce la emisión de otro más, tenemos *n+1* fotones con la misma fase, dirección, frecuencia y polarización, con lo que la intensidad de la radiación formada por esos fotones aumenta. Si somos capaces de producir este tipo de emisión en una longitud *d* de materia podremos tener más fotones *idénticos* a la salida que en la entrada: podemos "amplificar" la señal luminosa. Podemos pues ver que este tipo de radiación va a ser, precisamente, de capital importancia en el funcionamiento de los amplificadores de luz y láseres. Ahora, la dinámica de los niveles ocupados por electrones inducida por este proceso vendrá dada por:

Emisión estimulada de un fotón
Es un proceso por el cual un fotón origina una transición de un electrón de un nivel de mayor energía a otro de menor. El resultado es que el fotón saliente tiene todas sus propiedades idénticas al que induce la emisión, con lo que el proceso termina con dos fotones. De esta forma se puede obtener mayor número de fotones idénticos a la salida que a la entrada: Se puede "amplificar" la intensidad de la radiación.

$$\left.\frac{dN_2}{dt}\right)_{em\ est} = -B_{21}\cdot\rho(\nu)\cdot N_2 \quad . \tag{8.10}$$

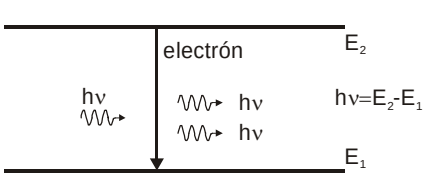


Figura 8.4.3 La presencia de un fotón provoca *(estimula)* la transición de un electrón del nivel 2 al 1. El fotón emitido en esta transición es idéntico al que la origina, y el proceso acaba con dos fotones idénticos

Por lo tanto, la variación total de la densidad de niveles 2 ocupados por electrones viene dada por la suma de los distintos procesos, y es:

$$\boxed{\frac{dN_2}{dt} = -A_{21} \cdot N_2 + B_{12} \cdot \rho(\nu) \cdot N_1 - B_{21} \cdot \rho(\nu) \cdot N_2}. \qquad (8.11)$$

Esta ecuación recibe el nombre de ecuación de Einstein, y constituye el punto de partida para el estudio de todos los procesos optoelectrónicos.

Los coeficientes de proporcionalidad A_{21}, B_{21} y B_{12} están relacionados entre sí. Esta relación se puede obtener imponiendo que en equilibrio termodinámico N_1 y N_2 satisfagan la relación de Boltzmann y que $\rho(\nu)$ sea la del cuerpo negro:

$$\frac{N_2}{N_1} = \frac{g_2}{g_1} \cdot e^{-\frac{E_2 - E_1}{k_B T}} \; ; \quad \rho(\nu) = \frac{8\pi n^2 n_g \nu^2}{c^3} \cdot \frac{h\nu}{e^{\frac{h\nu}{k_B T}} - 1} \cdot \qquad (8.12)$$

g_1 y g_2 son la degeneración (número de electrones que pueden ocupar cada nivel atómico) del nivel 1 y 2 respectivamente, y k_B la constante de Boltzmann. Así se obtiene:

$$\boxed{g_2 B_{21} = g_1 B_{12}; \quad B_{21} = A_{21} \frac{c^3}{8\pi n^2 n_g h \nu^3}}. \qquad (8.13)$$

Consideremos ahora una situación en la que únicamente tengamos emisión espontánea: Por ejemplo, iluminamos un material, lo que provoca transiciones desde los niveles de menor energía a los de mayor, con lo que los estados más energéticos son ocupados por electrones. Si apagamos la luz, estos electrones permanecen un cierto tiempo en el nivel superior, y su caída al inferior se produce por emisión espontánea ya que $\rho(\nu)$ es despreciable. Sea N_2 la densidad de niveles altos ocupados por electrones. Su evolución temporal tras apagar la luz la podemos obtener a partir de la ecuación de Einstein, que tendrá la forma

$$\frac{dN_2}{dt} = -A_{21} N_2 = -\frac{N_2}{\tau_{esp}} \quad \text{si definimos} \quad \tau_{esp} = \frac{1}{A_{21}}. \qquad (8.14)$$

Integrando, obtenemos:

$$N_2(t) = N_{20} \cdot e^{-\frac{t}{\tau_{esp}}}. \qquad (8.15)$$

El número de fotones que se emiten por unidad de tiempo es precisamente la derivada de $N_2(t)$ con respecto al tiempo. Estos fotones tienen una energía $h\nu$ cada uno por lo que la energía emitida por unidad de tiempo, es decir, la potencia, será:

$$W = h\nu \cdot \frac{dN_2}{dt} = h\nu \frac{N_{20}}{\tau_{esp}} e^{-\frac{t}{\tau_{esp}}} . \tag{8.16}$$

Los fotones salen en todas direcciones. Si nuestro detector determina un ángulo sólido Ω, los que detectará será $\Omega/(4\pi)$ del total. Considerando que este ángulo sólido corresponda a un área A, la intensidad que incidirá sobre ella será:

$$I = \frac{\Omega}{4\pi}\frac{W}{A} = \frac{\Omega}{4\pi A} \cdot h\nu \cdot \frac{dN_2}{dt} = \frac{\Omega}{4\pi A} \cdot h\nu \frac{N_{20}}{\tau_{esp}} e^{-\frac{t}{\tau_{esp}}} . \tag{8.17}$$

Lo que nos indica que la intensidad medida será proporcional a una exponencial de constante de tiempo τ_{esp}, que representa justamente *la vida media espontánea* del nivel 2, y por eso recibe este nombre. En lo sucesivo utilizaremos indistintamente A_{21} o τ_{esp}, dependiendo de lo que sea más habitual en la práctica según el contexto.

8.5 Coeficiente de ganancia de intensidad de la radiación luminosa

Vamos a dar ahora una definición matemática de la ganancia de intensidad de una radiación luminosa, sin proceder a una demostración rigurosa de la misma. La ganancia de intensidad está asociada a los mecanismos de absorción y emisión estimulados que dependen de la intensidad de la radiación presente a través de $\rho(\nu)$. Como ya se ha indicado, cuanto mayor sea esta intensidad mayor serán el número de fotones que se añadirán (y absorberán) en un trozo de material. Además, si tenemos en cuenta que el fotón estimulado es idéntico al que estimula, aquel sale en su misma dirección. Así, podemos considerar un haz de fotones idénticos (misma energía o frecuencia, fase, dirección y polarización) que se propagan en la dirección z constituyendo un haz monocromático de intensidad I_ν, y un trozo de material de longitud dz en dicha dirección (Figura 8.5.1). La intensidad que saldrá del mismo será $I_\nu + dI_\nu$. El incremento dI_ν *depende* de I_ν (intensidad presente en el inicio de dz), por lo que en cada trozo dz la producción de fotones será mayor que en el anterior, en la mayoría de los casos de forma no lineal. Teniendo en cuenta esto, se define el *coeficiente de ganancia de intensidad de radiación luminosa* $\gamma(\nu)$ como el incremento relativo de intensidad por unidad de longitud:

$$\boxed{\gamma(\nu) = \frac{1}{I_\nu}\frac{dI_\nu}{dz}} \tag{8.18}$$

La Ganancia de intensidad de radiación luminosa es el incremento relativo de intensidad por unidad de longitud.

γ depende del material y en muchos casos de la posición dentro del material, esto es $\gamma = \gamma(\nu, z, \text{propiedades del material})$. Si $\gamma > 0$ la derivada de (8.18) es positiva, lo que significa que la intensidad aumenta.

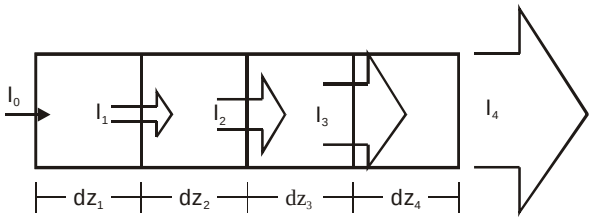


Figura 8.5.1 La intensidad que sale de cada elemento *dz* depende de la que entra en el mismo.

El coefiente de absorción es la disminución relativa de intensidad por unidad de longitud.

La relación entre ellos es $\alpha = -\gamma$

Cuando $\gamma < 0$, se suele reemplazar $|\gamma|$ por α, denominado *coeficiente de absorción del material* y la expresión anterior queda:

$$-\alpha = \frac{1}{I_\nu}\frac{dI_\nu}{dz}. \tag{8.19}$$

Y si α es constante en todo el material, (8.19) se integra inmediatamente y sale:

$$I_\nu(z) = I_0 e^{-\alpha z}, \tag{8.20}$$

en donde I_0 es la intensidad que incide sobre el origen del material que estamos considerando. (8.20) es la conocida fórmula que da la absorción de luz en un material. Tanto γ como α dependen, en general, de la frecuencia de la intensidad luminosa que estimula las transiciones.

Otra forma muy utilizada de expresar la ganancia se obtiene multiplicando el numerador y denominador del segundo miembro de (8.18) por el área *A* sobre la que incide I_ν:

$$\boxed{\gamma = \frac{\dfrac{d(A\cdot I_\nu)}{A\cdot dz}}{I_\nu} = \frac{\dfrac{\text{Incremento de potencia generada}}{\text{Volumen en el que se genera la potencia}}}{\text{Intensidad incidente}}}. \tag{8.21}$$

O, vuelta a escribir de otra forma equivalente:

$$\boxed{\gamma = \frac{\text{Incremento de potencia generada por unidad de volumen}}{\text{Intensidad incidente}}}. \tag{8.22}$$

Podemos aún relacionar γ con el número de fotones, dividiendo el numerador y el denominador de la expresión anterior por hν. Y así, tenemos:

$$\boxed{\gamma = \frac{\text{Incremento del número de fotones generados por unidad de volumen}}{\text{Número de fotones incidentes por unidad de área}}} \tag{8.23}$$

8.6 Revisión de la teoría elemental de semiconductores

La única forma de explicar correctamente las propiedades eléctricas y ópticas de los sólidos es utilizando la mecánica cuántica, lo que introduce una dificultad tanto conceptual como matemática muy grande. Para que en nuestro nivel podamos hacer un uso adecuado de la misma es preciso admitir bastantes ideas sin demostración y hacer simplificaciones muy considerables, con lo que podremos reducir notablemente la complejidad. Aún así los resultados que se obtienen tienen un grado de aproximación bastante bueno y nos permitirán obtener valores razonables de las propiedades que nos interesen, en el sentido de que muestran una buena coincidencia con los datos obtenidos experimentalmente. Aunque en el capítulo 1 se ha dado una visión cualitativa de las características de los semiconductores, vamos ahora a profundizar un poco más, especialmente en lo que concierne a la estructura de bandas, ya que nos va a resultar imprescindible para explicar los procesos de interacción entre electrones y fotones en un material de este tipo.

Para empezar, lo primero que vamos a admitir es que los sólidos que nos interesan son sólidos cristalinos, esto es, que los átomos que los constituyen están dispuestos periódicamente en el espacio, ocupando posiciones fijas en el espacio. Este modelo presenta algunas dificultades conceptuales, ya que los núcleos y los electrones más cercanos a ellos, que no participan de la conducción, también se mueven. No obstante, no tendremos en cuenta esta situación, considerándolos fijos. Esta será, pues, la primera de las aproximaciones citadas, (llamada, en este contexto aproximación adiabática).

Lo característico de la distribución periódica de los átomos es que tomando como origen uno cualquiera de ellos, los demás están dispuestos en puntos "discretos" $\mathbf{R}$ tales que

$$\mathbf{R} = n_1\mathbf{a_1} + n_2\mathbf{a_2} + n_3\mathbf{a_3} \qquad (8.24)$$

siendo n_1, n_2 y n_3 números enteros, y $\mathbf{a_1}$, $\mathbf{a_2}$ y $\mathbf{a_3}$ los tres vectores más cortos que unen el origen con sus tres vecinos más próximos. El conjunto de puntos $\mathbf{R}$ se denomina *red de Bravais*, y, básicamente, es una idealización matemática del cristal, ya que, en muchos casos, no basta con asociar a un punto de la red un sólo átomo, sino que hay que asociarle un grupo de átomos. Este grupo constituye la *base* del cristal, de forma que

Sólido cristalino = red de Bravais + base

Con estas aproximaciones, se puede demostrar que:

- Los electrones del cristal se pueden tratar como independientes, de forma que cada uno de ellos se puede describir mediante una función de onda solución de la ecuación de Schrödinger :

$$-\frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{r})+V_0(\mathbf{r})\psi(\mathbf{r})=E\,\psi(\mathbf{r}), \qquad (8.25)$$

donde $V_0(\mathbf{r})$ es el potencial periódico de la red al que está sometido el electrón considerado y que cumple la condición de periodicidad $V_0(\mathbf{r}+\mathbf{R}) = V_0(\mathbf{r})$.

- Las soluciones de la ecuación anterior han de ser de la forma

$$\psi_{\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}\,u_{\mathbf{k}}(\mathbf{r}); \quad \text{con} \quad u_{\mathbf{k}}(\mathbf{r}+\mathbf{R}) = u_{\mathbf{k}}(\mathbf{r}), \qquad (8.26)$$
$$E=E(\mathbf{k}); \quad E(\mathbf{k}) = E(-\mathbf{k}).$$

es decir, las soluciones dependen de un vector **k** que juega el papel de número cuántico. El resultado expresado en la primera línea de (8.26) recibe el nombre de Teorema de Bloch, y las funciones $\psi(\mathbf{r})$ reciben el nombre de funciones de Bloch. Para cada **k**, la función u(**r**) es diferente, lo que hemos indicado por el subíndice k. Por la misma razón, lo ponemos también en $\psi(\mathbf{r})$. Si ahora reemplazamos $\psi_{\mathbf{k}}(\mathbf{r})$ en (8.25) por su expresión en función de $u_{\mathbf{k}}(\mathbf{r})$, la ecuación que ésta ha de cumplir es:

$$-\frac{\hbar^2}{2m}\left\{\nabla^2 u_{\mathbf{k}}(\mathbf{r})+\left[2i\mathbf{k}\cdot\nabla u_{\mathbf{k}}(\mathbf{r})\right]-k^2 u_{\mathbf{k}}(\mathbf{r})\right\}+V_0(\mathbf{r})\,u_{\mathbf{k}}(\mathbf{r})=E(\mathbf{k})u_{\mathbf{k}}(\mathbf{r}) \qquad (8.27)$$

Si fijamos el valor de **k**, esta ecuación es del tipo de las de valores propios, que, en general, tiene infinitas soluciones, por lo que necesitaremos otro número para señalar cada una de estas. Llamemos *n* a este otro número, y los distintos valores de $\psi_{\mathbf{k}}(\mathbf{r})$, $u_{\mathbf{k}}(\mathbf{r})$ y $E(\mathbf{k})$ los tendremos que señalar como $\psi_{n,\mathbf{k}}(\mathbf{r})$, $u_{n,\mathbf{k}}(\mathbf{r})$ y $E_n(\mathbf{k})$. En la Figura 8.6.1 se muestra un diagrama de los posibles valores de la energía para todo **k** en una dimensión y distintos valores consecutivos de *n*.

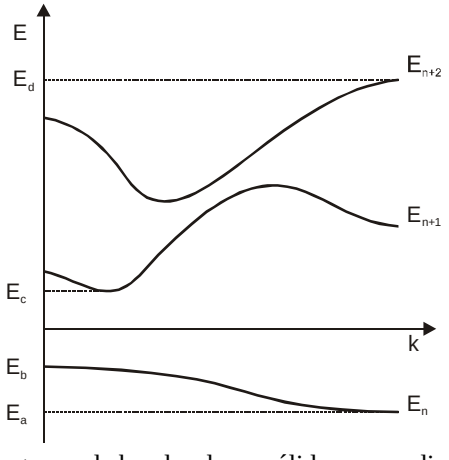


Figura 8.6.1 Diagrama de bandas de un sólido en una dimensión. Se muestran tres bandas, E_n, E_{n+1} y E_{n+2}

Capítulo 8

Bandas de Energía
Los materiales que tienen una estructura cristalina también tiene una estructura de bandas.

Esto no es privativo de materiales cristralinos. Los que no lo son también pueden tener estructura de bandas

En ella se puede ver que un electrón puede tener cualquier valor de la energía entre E_a y E_b, y entre E_c y E_d. En cambio no hay ningún valor de k que permita los valores entre E_b y E_c. Las zonas de energía permitida se llaman bandas permitidas, y las que no son posibles, bandas prohibidas, o "gaps" (utilizando la terminología inglesa). No obstante, en la práctica se confunde esta definición con la expresión $E_n(k)$. Así, si decimos que "el electrón está en la tercera banda con k_0 ", queremos expresar que el electrón tiene el valor de la energía que corresponde a $E_3(k_0)$.

- Como $E_n(\mathbf{k}) = E_n(-\mathbf{k})$ en $\mathbf{k} = 0$ tiene que haber un máximo o un mínimo relativo (un extremo relativo) de la energía.
- En un semiconductor, a T = 0 K, están ocupadas todas las bandas hasta la que determina el límite inferior de la banda prohibida o gap. (Banda E_n en la Figura 8.6.1). Las bandas superiores están completamente vacías. A temperatura ambiente, algunos de los electrones de la última banda pasan, adquiriendo energía de la red cristalina, a la banda superior, quedando estados no ocupados -huecos- en la banda inferior y estados ocupados por electrones en la superior. La última banda que está ocupada a T=0 K se denomina banda de valencia, y la primera vacía, banda de conducción.
- Los valores que puede tomar **k** dependen de los vectores $\mathbf{a_1}$, $\mathbf{a_2}$ y $\mathbf{a_3}$ que elijamos, pero para lo que necesitaremos podemos admitir un cristal en el que tales vectores estén sobre los ejes de coordenadas de forma que si las dimensiones del cristal son L_x , L_y y L_z , y la separación entre los puntos de la red de Bravais (en cada punto puede haber más de un átomo si la base es poliatómica) es a_x , a_y y a_z , el número de puntos de dicha red de Bravais en cada dirección habrá de ser $N_i = L_i/a_i$. (i=x,y,z). Pues bien, cada componente del vector **k**, k_i, debe valer:

$$k_i = \frac{2\pi}{N_i a_i} m_i \qquad (i = x, y, z); \quad m_i \text{ es un número entero} \qquad (8.28)$$

(Para una deducción rigurosa más general ver las referencias 0 y 0).

En los materiales de interés para nosotros, $a_x = a_y = a_z = a$, y se denomina *constante de red*. Para un material típico en optoelectrónica, el arseniuro de galio, GaAs, a = 0.565 nm (= 5.65 Å).

- Los valores de m_i mayores que N_i, si bien son posibles, no aportan ninguna información ya que se puede demostrar que las propiedades físicas, eléctricas y ópticas de m_i y de $m_i - N_i$ son las misma. Por lo tanto,

m_i es un número entero que vale $\quad 0 \le m_i < N_i \quad (i = x, y, z)$ (8.29)

El vector de onda **k** tiene valores discretos, y hay tantos valores permitidos como puntos de la red de Bravais tiene el cristal.

Por lo tanto, hay N_i posibles valores para cada componente de **k**. En cada banda de energía hay N_1 x N_2 x N_3 =N estados posibles, esto es tantos estados como puntos de la red de Bravais del cristal. Si tenemos en cuenta que los sólidos tienen del orden de 10^{22} átomos por centímetro cúbico, podemos estimar $N_i \approx (10^{22})^{1/3} = 2.15 \times 10^7$ (considerando que en cada punto de la red de Bravais hay un átomo) para un cristal cúbico de *1 cm* de lado. Incluso para un cristal cúbico de *1 μm* de lado ($=10^{-12}$ *cm*^{3}) el número de átomos sería:

$$N_i \approx (10^{22}\ at/cm^3 \times 10^{-12}\ cm^3)^{1/3} = 2.15 \times 10^3\ átomos$$

que representa un número bastante grande (si dibujamos el eje k_x por ejemplo, sobre una longitud de *1m*, la separación entre valores de k_x sería inferior a *0.5 mm*). Esto significa que hay muchos valores de k_i posibles en cada dirección.

La separación en el espacio de las **k** entre cada uno de sus valores la podemos obtener teniendo en cuenta que entre k_i y k_{i+1} hay una distancia Δk_i

$$\Delta k_i = k_{i+1} - k_i = \frac{2\pi}{N_i a_i} = \frac{2\pi}{L_i} \quad (i = x, y, z) \tag{8.30}$$

por lo que podemos considerar que cada **k** (cada estado electrónico) "ocupa" un volumen $\Delta\mathbf{k}$ dado por:

$$\boxed{\Delta\mathbf{k} = \frac{2\pi}{Lx} \cdot \frac{2\pi}{L_y} \cdot \frac{2\pi}{L_z} = \frac{(2\pi)^3}{V_C}}, \tag{8.31}$$

siendo V_C el volumen del cristal. Este resultado, aunque deducido para un caso muy particular, vale de forma general.

- El vector de onda **k** desempeña el papel que en mecánica clásica tiene el momento lineal con respecto a la conservación del mismo: En las interacciones del electrón con fotones o con vibraciones de la red (denominadas *fonones*), el vector de onda total se conserva. Así, si un fotón tiene un vector $\mathbf{k_{ph}}$ y es absorbido por un electrón con un vector de onda $\mathbf{k}_i$, el electrón acabará con un vector de onda $\mathbf{k_f}$ tal que:

$$\mathbf{k}_f = \mathbf{k}_i + \mathbf{k}_{ph}\,. \tag{8.32}$$

O, si el proceso fuera de emisión del fotón por parte del electrón:

$$\mathbf{k}_f = \mathbf{k}_i - \mathbf{k}_{ph}\,. \tag{8.33}$$

k recibe el nombre de *pseudomomento* o *momento cristalino* del electrón, ya que no es momento lineal total del electrón, pero tiene las propiedades de conservación mencionadas.

Supongamos que un electrón pasa de un estado en una banda E_{n+1} a otro estado en otra banda de menor energía E_n emitiendo un fotón de $\lambda = 500\ nm$ (que corresponde al color amarillo - verdoso, en la zona visible del espectro). El fotón emitido tiene un vector de onda (considerando $n = 1$):

$$k_{ph} = \frac{2\pi}{\lambda} = \frac{2\pi}{5} \cdot 10^5\ cm^{-1} = 2\pi \cdot 2 \cdot 10^4\ cm^{-1}$$

Tomemos el valor de a para este material de $0.5\ nm$, por lo que el máximo valor de k que puede tener el electrón es k_m

$$k_m = \frac{2\pi}{a} = \frac{2\pi}{0.5} \cdot 10^7\ cm^{-1} = 2\pi \cdot 2 \cdot 10^7\ cm^{-1} = 10^3 \cdot k_{ph}$$

Si ahora dibujamos a escala en un diagrama de bandas esta transición, (Figura 8.6.2), partiendo de un valor para k_i nos encontraremos con que el valor de k_f, que cambia en una milésima de k_m, tenemos que dibujarlo vertical, ya que no disponemos de tanta resolución gráfica como necesitaríamos para ver esta variación. Esto es cierto no sólo gráficamente, sino también con la precisión con que podemos calcular los valores k_i y k_f, y la condición que normalmente se toma en las transiciones ópticas es

$$\mathbf{k}_f = \mathbf{k}_i$$

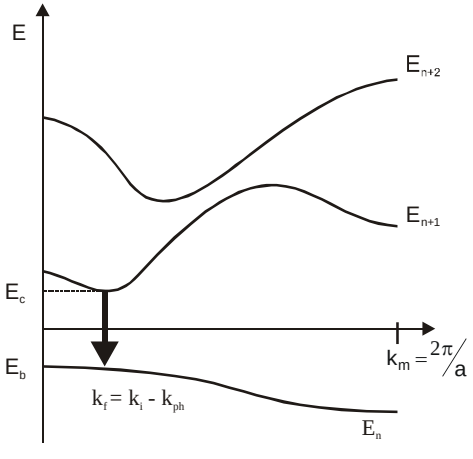


Figura 8.6.2 Diagrama de bandas de un sólido mostrando una transición con emisión de un fotón entre dos bandas. Obsérvese que la representación de la gráfica ha de ser vertical

Esta condición es muy importante, y nos explica porqué unos materiales tienen buenas propiedades optoelectrónicas y otros

no. Consideremos el caso de emisión de luz, y veamos en primer lugar porque el silicio cristalino no sirve como dispositivo emisor. Su banda de valencia tiene un máximo en **k** = 0, y la de conducción un mínimo en la dirección k_x (y también en k_y y k_z) aproximadamente a $0.85k_m$. A temperaturas normales de funcionamiento, los estados ocupados por electrones se sitúan entorno del mínimo de la banda de conducción, y los ocupados por huecos (vacíos de electrones) entorno del máximo de la banda de valencia, por lo que sólo se pueden dar transiciones entre estos extremos (los electrones no pueden realizar transiciones "verticales" desde el mínimo de la banda de conducción a la de valencia porque los estados de esta banda a los que deberían ir a parar los electrones provenientes de la de conducción ya están ocupados). Sin embargo, los fotones no tienen el vector **k** suficientemente grande para compensar la variación del momento cristalino del electrón y esta transición no puede tener lugar. En consecuencia, el silicio no emite luz. En cambio, el GaAs, tiene el máximo de la banda de valencia, y el mínimo de la banda de conducción en **k** = 0. La transición "vertical" es posible, y este material es uno de los más usados en optoelectrónica.

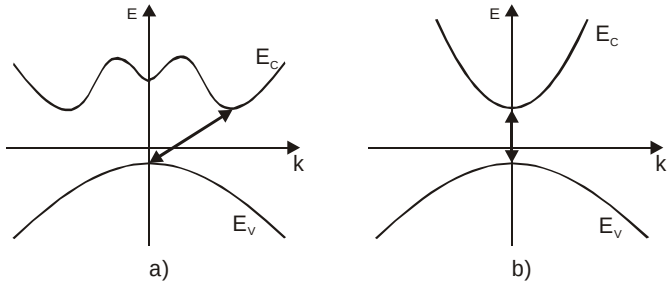


Figura 8.6.3 Diagrama esquemático de las bandas de conducción y valencia de un sólido a) de gap indirecto tipo silicio y b) de gap directo tipo GaAs. En el caso a) un fotón no puede producir una variación de k tan grande, por lo que no puede darse la transición indicada.

- En las situaciones normales de funcionamiento, se admite que los electrones en la banda de conducción y los huecos en la banda de valencia están situados en torno a un mínimo o máximo (ambos en el mismo valor de **k**). (Ver Capítulo 2). Si llamamos E_{C0} y E_{V0} al mínimo y máximo de la banda de conducción y valencia respectivamente, y m_c^*, m_v^* a las masas efectivas de electrones y huecos, la energía de los electrones en la banda de conducción la podemos expresar por:

$$E_C = E_{C0} + \frac{\hbar^2 k^2}{2 m_c^*}, \tag{8.34}$$

y la de los huecos en la de valencia por:

$$E_V = E_{V0} - \frac{\hbar^2 k^2}{2 m_v^*}. \tag{8.35}$$

Estas relaciones son una aproximación un poco burda a la forma de $E_{C,V}(\mathbf{k})$, muy especialmente en el caso de la banda de valencia, en la que, incluso los modelos más simples obligan a considerar al menos dos bandas de diferente curvatura, las denominadas banda de huecos ligeros y banda de huecos pesados. Aún así, nos quedaremos con una sola banda para poder establecer con cierta simplicidad las ideas básicas y obtener expresiones analíticas que describen bastante bien los resultados experimentales.

Finalmente, la probabilidad de que un estado de energía E_1 esté ocupado por un electrón viene dada, en equilibrio termodinámico, por la función de Fermi-Dirac:

$$f(E) = \frac{1}{1 + e^{\frac{E - E_F}{k_B T}}}. \tag{8.36}$$

T es la temperatura absoluta, k_B la constante de Boltzmann y E_F el nivel de Fermi (o potencial químico) y es independiente de la posición. Cuando no estamos en equilibrio, utilizaremos los *pseudoniveles de Fermi*, E_{Fn} y E_{Fp} y la probabilidad de que un estado de la banda de conducción esté *ocupado por electrones* vendrá dado por

$$f_c(E) = \frac{1}{1 + e^{\frac{E - E_{Fn}}{k_B T}}}, \tag{8.37}$$

y la de que un estado en la *banda de valencia esté ocupado por electrones por*

$$f_v(E) = \frac{1}{1 + e^{\frac{E - E_{Fp}}{k_B T}}}. \tag{8.38}$$

En equilibrio termodinámico, $E_{Fn} = E_{Fp} = E_F$ y las expresiones anteriores se reducen a (8.36).

8.7 Materiales semiconductores utilizados en dispositivos optoelectrónicos

Los materiales utilizados en dispositivos optoelectrónicos son muy variados, y su uso depende de la aplicación y rango de frecuencias a que vayan destinados. En las zonas del espectro correspondientes al infrarrojo cercano y al visible se usan aleaciones de materiales de la

columna III y V del sistema periódico, tipo GaAs (Arseniuro de Galio), AlGaAs (Arseniuro de Galio y Aluminio), GaN (Nitruro de Galio), GaP (Fosfuro de Galio), InP (Fosfuro de Indio), y mezclas entre ellos, y en el infrarrojo medio o lejano algunos de estos así como compuestos de la columna II y IV (CdHgTe, PbSSe...), casi siempre materiales de gap directo. La razón ya ha sido mencionada. Los de gap indirecto pueden jugar también algún papel si se impurifican convenientemente, de forma que se creen niveles permitidos en la banda prohibida que pueden ayudar a modificar el valor de **k** en una transición que vaya de una banda al nivel creado, y de allí a la otra banda. Estos procesos, por su complejidad, no los trataremos aquí.

El primer láser semiconductor que emitió luz fue de GaAs, y la combinación de este material con AlGaAs ha sido el más utilizado. Vamos, a modo de ejemplo, a ver principalmente este tipo de estructura. En la Figura 8.7.1 se muestra el ancho de banda en función de la constante de red. Como ejemplo típico, podemos ver que el GaAs y el AlAs tienen prácticamente la misma constante de red y diferente ancho de banda prohibida.

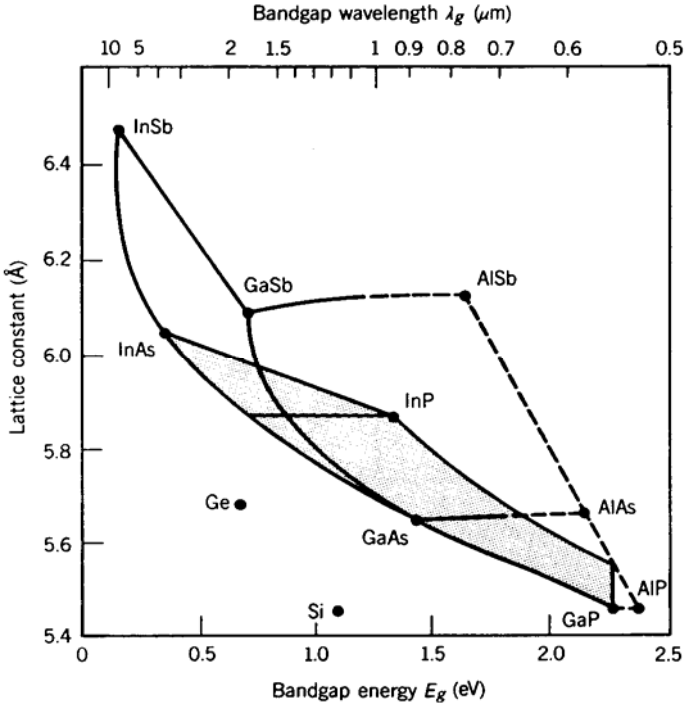


Figura 8.7.1 Ancho de la banda prohibida en función de la constante de red. Las líneas discontinuas muestran las combinaciones de compuestos en las que el gap es indirecto

Si ahora mezclamos adecuadamente GaAs y AlAs, de forma que reemplacemos un cierto número de átomos de Ga por átomos de Al

podemos obtener un compuesto denominado $Al_xGa_{1-x}As$, donde la x indica la proporción de átomos de Ga sustituidos por los de Al. Así, por ejemplo, $Al_{0.2}Ga_{0.8}As$ significa que un 20% de átomos de Ga han sido reemplazados por los de Al. El gap de este nuevo material está entra los valores de los del GaAs y AlAs, y en el caso citado es de 1.67 eV. Así pues, podemos variar el gap a voluntad (dentro de unos límites), y obtener materiales que pueden emitir o absorber radiación luminosa con una frecuencia umbral determinada. Además, en este ejemplo, la separación interatómica del nuevo material es muy aproximadamente la misma que la del GaAs, y eso da una nueva posibilidad: Se pueden crecer capas de AlGaAs sobre capas de GaAs sin que se produzcan tensiones que deformen la red, y por lo tanto, no cambien la estructura de bandas de los mismos. De ello resultan compuestos formado por capas de GaAs, con un gap de 1.42 eV, y capas de $Al_xGa_{1-x}As$ con otro gap mayor, por ejemplo de 1.67 eV. Estas estructuras compuestas presentan muchas ventajas, y de hecho constituyen la base de la optoelectrónica actual. A modo de ejemplo, podemos ver que en un dispositivo emisor de luz, si se consigue que la emisión se produzca en la capa de GaAs, las capas de AlGaAs no absorberán fotones porque tienen mayor gap, con lo que no disminuirán el rendimiento. Volveremos sobre ellas cuando hablemos de los láseres de heterounión. Lo que hemos referido, a modo de ejemplo al par GaAs y AlAs vale para otros pares como GaP-AlP y (GaAs-InAs) - InP, dando, en este caso, compuestos cuaternarios, esto es, formados por cuatro elementos. En la Tabla 1 se dan algunos datos para los materiales más habituales:

Material	Ancho de banda prohibida (eV)	Constante de red (nm)	Constante dieléctrica relativa - ε_r
GaAs	1.43	0.5653	13.2
AlAs (i)	2.16	0.5660	10.9
GaP (i)	2.26	0.5451	11.1
InP	1.35	0.5969	12.4

Tabla 8.7.1 Propiedades más significativas de algunos de los materiales usados en optoelectrónica. El símbolo (i) indica que la banda prohibida es indirecta.

8.8 Descripciones cualitativas de la absorción y de la ganancia de luz en un semiconductor

Con el fin de empezar a entender cómo son los procesos optoelectrónicos en un semiconductor, consideremos algunas situaciones ideales que nos servirán para ir describiendo cualitativamente las

transiciones entre bandas que en ellos tienen lugar, y fijar algunos conceptos que luego utilizaremos de forma cuantitativa.

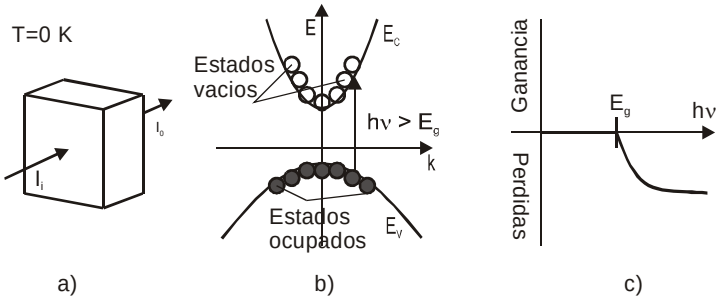


Figura 8.8.1 En a) se ilumina una muestra semiconductora que se mantiene a 0 K. b) Diagrama de bandas, mostrando la ocupación a 0 K. c) Forma de la ganancia o perdida de luz del semiconductor.

En primer lugar, supongamos que tenemos un semiconductor intrínseco a T = 0 K. En este caso, los estados de la banda de valencia estarán completamente ocupados por electrones, y los de la de conducción, totalmente vacíos (esta es la característica que define a un semiconductor). Supongamos ahora que iluminamos el semiconductor con un haz de luz monocromático cuya frecuencia podemos variar y que no sea muy intenso (Figura 8.8.1 a). Si la frecuencia de la radiación es tal que su energía $h\nu$ es menor que la de la banda prohibida, E_g, entonces no hay transiciones posibles ya que si bien hay muchos electrones en la banda de valencia, no hay estados adonde puedan ir. Por lo tanto, si prescindimos de las reflexiones que se producen en las superficies que limitan al semiconductor, y que son inherentes a cualquier separación de medios de propagación, la intensidad incidente sería igual a la transmitida. No se absorberían fotones, por lo que el material sería transparente (Figura 8.8.1 b). Cuando la energía de la luz supere la de la banda prohibida, los electrones en la banda de valencia podrán saltar a la de conducción, absorbiendo fotones. La intensidad de la luz saliente I_0 será menor que la de la luz incidente I_i. Si definimos la "Ganancia" como:

$$Ganancia = \log (I_0 / I_i) = - Perdidas$$

este parámetro tiene la forma que se muestra en la Figura 8.8.1 c). En esta situación no puede haber ganancia, tan sólo pérdida de luz.

Obsérvese que al describir la anterior situación hemos considerado que la radiación incidente no fuera muy intensa, y, aunque parezca que esto no nos ha servido para nada, no es así. Siempre que hay radiación, hay una parte de la misma que se absorbe: hay estimulación. Son dos procesos que se manifiestan juntos. Por lo tanto, en la explicación anterior deberíamos haber añadido un término de emisión estimulada. Si

la radiación incidente no es muy intensa, generará "pocas" transiciones (por unidad de tiempo) y la concentración de electrones en la banda de conducción nunca será comparable a la que hay en la banda de valencia, por lo que la contribución de la emisión estimulada es despreciable.

Supongamos ahora que, manteniendo la temperatura a 0 K, y el haz de prueba a frecuencia variable, iluminamos uniformemente el semiconductor con una luz de energía $h\nu_1 > E_g$, de forma que hay electrones ocupando estados en la banda de conducción y estados vacíos en la de valencia. Como estamos a 0 K, la banda de conducción estará llena hasta E_{Fn} y la de valencia hasta E_{Fp} (Figura 8.8.2).

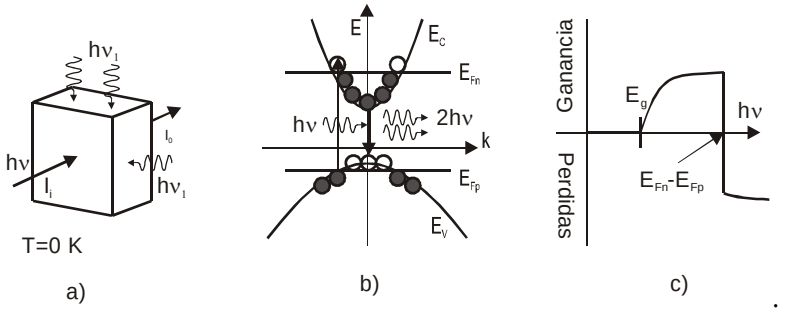


Figura 8.8.2 En a) se ilumina una muestra semiconductora con una luz de frecuencia $h\nu_1$ además de la de test $h\nu$. La muestra se mantiene a 0 K. b) Diagrama de bandas, mostrando la ocupación a 0 K la presencia de $h\nu_1$ causa que se ocupen estados en la banda de conducción y se vacíen en la de valencia . c) Forma de la ganancia o perdida de luz del semiconductor.

Con esta nueva situación, los procesos cambian radicalmente. Ahora, un fotón de energía ligeramente superior a E_g no puede ser absorbido, ya que los estados con esta diferencia de energía de la banda de valencia están vacíos y los de la de conducción ocupados. En cambio, si pueden emitir por fotones estimulados en el mismo estado que el que lo provoca: Podemos tener amplificación. A 0 K, esto se puede dar en el intervalo de frecuencias:

$$E_g < h\nu < E_{Fn} - E_{Fp}\,. \tag{8.39}$$

O sea, en este intervalo podemos tener ganancia, si conseguimos que haya suficientes electrones en la banda de conducción y huecos en la de valencia. Para energías superiores a E_{Fn}-E_{Fp} de nuevo tendríamos absorción sin estimulación (Este proceso está representado en la Figura 8.8.2 b) mediante una flecha que va desde la banda de valencia a la de conducción). La forma de la curva de ganancia es la que se muestra en la Figura 8.8.2 c), y es justamente la misma que la de la Figura 8.8.1 c) pero invertida. La emisión y la absorción son fenómenos inversos uno del otro.

Si la temperatura no es de 0 K, la separación entre estados vacíos y ocupados por electrones en ambas bandas no está bien definida. Aún

así, la forma de la curva es muy parecida a la dibujada, con los bordes de la línea vertical redondeados: El paso de ganancia a perdidas ya no es brusco, sino continuo, aunque muy rápido, salvo que la temperatura sea muy elevada.

8.9 Amplificadores ópticos y Láseres semiconductores

De acuerdo con lo dicho, si conseguimos tener un número alto -más correctamente, una densidad elevada- de electrones en la banda de conducción y huecos en la de valencia podemos generar fotones mediante emisión estimulada y amplificar la luz. Podemos, pues, conseguir un *amplificador óptico*. Como veremos más adelante, esta amplificación puede tener lugar para todas las frecuencias que se den en el intervalo $G_E < h\nu < C_{FN}\text{-}F_{EPP}$. En realidad, no todas las transiciones de electrones de una banda hacia la otra emiten fotones. Muchas de esas transiciones ceden su energía a la red, calentando el semiconductor. La eficiencia de la transición óptica no es la misma a todas las frecuencias.

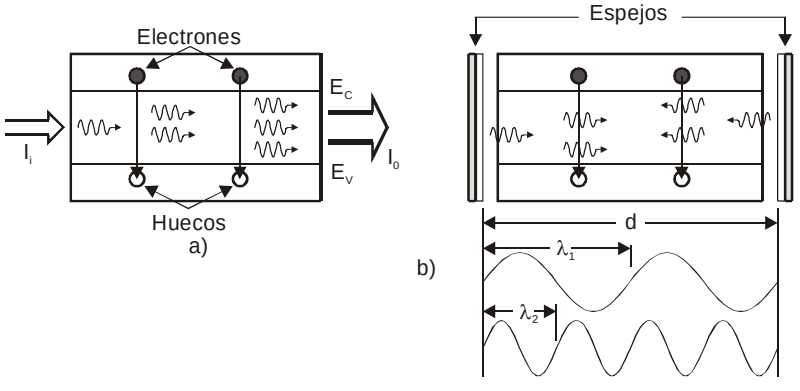


Figura 8.9.1 En a) se muestra un amplificador óptico. La luz que entra se amplifica. En b) se han añadido dos espejos. Ya no hace falta luz externa. La emisión espontánea inicia el proceso que se mantiene por la acción de los espejos.

Si ahora nosotros ponemos dos espejos en los extremos del material semiconductor, separados por una distancia *d*, de forma que, al reflejarse la luz entre ellos repetidamente, puedan formarse ondas estacionarias, entonces en la cavidad óptica delimitada por los espejos aumenta la densidad de energía correspondiente a las frecuencias que puedan cumplir la condición de ondas estacionarias:

$$m \cdot \frac{\lambda}{2} = d \quad \acute{o} \quad m \cdot \frac{c}{2\,d\,n} = \nu \quad (m \text{ es un número entero}). \qquad (8.40)$$

Cómo habrá más densidad de energía a estas frecuencias, se inducirán más transiciones estimuladas, lo que a su vez incrementará más

la densidad de energía, que producirá más emisiones estimuladas, etc. De esta forma tendremos amplificación a unas determinadas energías, es decir, un amplificador con selección de modos (cada par m, ν que cumplen la ecuación (8.40) recibe el nombre de modo). Pues bien. un amplificador óptico en el que podamos amplificar selectivamente uno o varios modos recibe el nombre de LÁSER.

En la práctica, los dispositivos LÁSER no requieren de ningún haz externo para su funcionamiento. Consideremos de nuevo la estructura, sin radiación externa, y con una elevada densidad de electrones en la banda de conducción y de huecos en la de valencia. Estos electrones y huecos se recombinarán espontáneamente, emitiendo fotones en todo el rango de frecuencias posible y en todas direcciones. La gran mayoría de estos fotones (por ejemplo, los que salgan en dirección paralela a los espejos) se perderán al salir del semiconductor o ser absorbido de nuevo por el mismo. Pero aquellos que salgan en la dirección de los espejos serán reflejados una y otra vez, induciendo, en cada paso entre los espejos, la emisión estimulada de nuevos fotones. De esta forma los modos estacionarios irán incrementando su número de fotones, es decir, su intensidad, hasta que se alcance una situación estacionaria definida por el hecho de que la velocidad a la que se recombinan los pares electrón-hueco ha de ser igual a la velocidad a la que desde el exterior se aportan pares electrón-hueco, ya que, en definitiva, no podemos obtener de ningún sistema más energía de la que aportamos.

Las siglas LÁSER son el acrónimo de **L**ight **A**mplificación by **E**mision of **S**timulated **R**adiation, en clara referencia al mecanismo físico que tiene lugar. Como ya hemos dicho, en la práctica hay que añadirle, además, la selección de frecuencia.

Podemos también ver un láser como un oscilador. En la Figura 8.9.2 se ha dibujado de forma muy esquemática un láser: Una pieza de material que proporciona una ganancia G de la intensidad luminosa, y dos

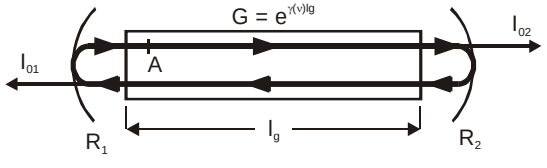


Figura 8.9.2 El Láser como oscilador. Se ha supuesto que $\gamma(\nu)$ no depende de la posición en el material activo. En casi todos los láseres reales, R_D o $R_D = 1$ con lo que H_{ilo} o I_{OA} es despreciable, y la luz sólo sale por un extremo

espejos con un coeficiente de reflexión R_D y R_D en los que se han incluido *todas* las otras pérdidas de intensidad que se puedan producir en el sistema: zonas en las que sólo hay absorción de luz, reflexiones en las superficies que puedan separar distintos materiales, etc. Consideremos un

punto cualquiera (por ejemplo, al A de la Figura 8.9.2) entre los espejos, y sea I_{OA} la intensidad de la luz en ese punto en un instante determinado. Después de una vuelta completa, la intensidad en el mismo punto H_{ILA} será:

$$I_{A1} = (G \cdot I_{A0}) \cdot R_2 \cdot G \cdot R_1 = G^2 \cdot R_1 \cdot R_2 \cdot I_{A0} . \qquad (8.41)$$

Ahora se pueden dar tres situaciones:

- $G^2 \cdot R_1 \cdot R_2 < 1$ Después de cada vuelta completa del haz de luz, la intensidad decrece, por lo que al cabo de un cierto número de vueltas "la luz se apaga". El sistema acabaría teniendo sólo la radiación debida a la emisión espontánea.
- $G^2 \cdot R_1 \cdot R_2 > 1$ Esta situación representa un incremento de la intensidad luminosa. Se puede dar al principio para iniciar la oscilación, pero no indefinidamente ya que tiene que haber algún mecanismo que limite el aporte de energía.
- $G^2 \cdot R_1 \cdot R_2 = 1$ Es esta la condición de mantenimiento estable de la intensidad, y es la que representa un funcionamiento continuo.

Estas tres condiciones se corresponden punto por punto con las de cualquier circuito oscilador electrónico (construido con transistores o amplificadores integrados, resistencias, bobinas y condensadores), y describen respectivamente las situaciones de apagado, inicio y mantenimiento.

Cuando se diseña un sistema láser, normalmente se busca que cumpla la condición $G^2 \cdot R_1 \cdot R_2 = 1$. Esta condición se denomina la *condición umbral* de funcionamiento, y es un problema bastante difícil ya que implica tener conocimientos de las propiedades de los materiales que intervienen (para obtener G) y de las ópticas asociadas al tipo de cavidad que realmente se va a tener según sea el dispositivo (para calcular R_1, R_2 y todas las demás pérdidas que hemos resumido en ellas). Nosotros nos limitaremos exclusivamente a tratar de obtener *G* en casos muy sencillos, pero suficientemente ilustrativos.

Como en general consideraremos que los materiales con los que tratamos tienen propiedades uniformes en el espacio (una vez más, para simplificar), *G* lo podemos obtener de $\gamma(\nu)$, ecuación (8.18):

$$\gamma(\nu) = \frac{1}{I_\nu} \frac{dI_\nu}{dz} . \qquad (8.42)$$

Si $\gamma(\nu)$ no depende de la posición, en una ida y vuelta del haz de luz (cruza dos veces l_g), la intensidad I_1 vale:

$$I_1 = I_0 \cdot e^{2\gamma l_g} \quad \Rightarrow \quad G = e^{2\gamma l_g} . \qquad (8.43)$$

Así pues, y más concretamente, nuestro objetivo será calcular $\gamma(\nu)$.

Finalmente, conviene recalcar un último aspecto de los amplificadores y láseres. Hemos dicho que para que se produzca la amplificación deberá haber una elevada densidad de electrones y huecos. Por lo tanto, tendremos de alguna manera que generarlos. El proceso por el cual se generan estos electrones y huecos recibe el nombre de *bombeo("pumping")*. En los primeros láseres de estado sólido -por ejemplo, los de rubí-, en los que los electrones realizaban transiciones entre varios estados (aunque la parte estimulada era tan sólo entre dos de ellos), este bombeo se realizaba mediante un destello de luz. En los láseres semiconductores, que se construyen con estructuras tipo uniones p-n del mismo material (homouniones) o de distintos materiales (heterouniones), el bombeo se realiza por inyección de corriente eléctrica, lo que es mucho más eficiente en cuanto a la relación entre energía suministrada - energía obtenida (esto es, rendimiento energético) y además, resulta fácil de variar, por lo que se puede, de forma relativamente sencilla, modular la salida del láser y transmitir así información. Estas son algunas de las ventajas de los láseres semiconductores sobre los de otro tipo. En cambio, entre sus desventajas podemos señalar la poca potencia que en general se obtiene así como el hecho de que el haz de luz tenga un ángulo de apertura bastante notable. Con todo, desde un punto de vista de aplicaciones electrónicas y especialmente en comunicaciones, las ventajas superan claramente a las desventajas .

8.10 Coeficiente de ganancia en un semiconductor

Vamos a obtener el coeficiente de ganancia $\gamma(\nu)$ para el caso de un semiconductor en el que las transiciones se realizasen entre la banda de conducción y la de valencia y que ambas fueran esféricas, es decir, que su relación $E(k)$ fuera de la forma dada en las ecuaciones (8.34) y (8.35) y que para más comodidad repetimos a continuación:

$$E_C = E_{C0} + \frac{\hbar^2 k^2}{2\, m_c^*}; \quad E_V = E_{V0} - \frac{\hbar^2 k^2}{2\, m_v^*}. \tag{8.44}$$

Sea E_2 un valor permitido de la energía de un electrón en la banda de conducción, y dE_2 un intervalo infinitesimal de energía entorno a E_2. Los electrones que realicen transiciones con emisión de fotones deberán conservar su vector de onda **k**, y ello nos dará el valor de la energía E_1 y el correspondiente intervalo dE_1 de la banda de valencia al que deberán ir a parar (ver Figura 8.10.1).

Llamemos $\mathbf{k_2}$ al valor de **k** cuya energía es E_2. Este mismo valor de **k** será el que corresponderá a E_1. Por lo tanto,

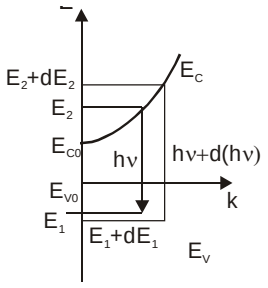


Figura 8.10.1 a) La forma de las bandas determina E_1 y dE_1 una vez elegidos E_2 y dE_2 o viceversa.

$$\left.\begin{aligned}\frac{\hbar^2 k_2^2}{2} &= m_c^*(E_2 - E_{C0})\\ \frac{\hbar^2 k_2^2}{2} &= m_v^*(E_{v0} - E_1)\end{aligned}\right\} \Rightarrow \begin{cases} m_c^*(E_2 - E_{C0}) = m_v^*(E_{v0} - E_1) \\ m_c^*\, dE_2 = -m_v^*\, dE_1 \end{cases}. \qquad (8.45)$$

La energía del fotón emitida en esta transición es:

$$h\nu = E_2 - E_1 = (E_2 - E_{C0}) + (E_{C0} - E_{V0}) + (E_{V0} - E_1), \qquad (8.46)$$

y como E_{C0} -E_{V0} = E_g,

$$h\nu = (E_2 - E_{C0}) + E_g + \frac{m_c^*}{m_v^*}(E_2 - E_{C0}) \quad \text{o, lo que es lo mismo}$$

$$E_2 - E_{C0} = \frac{m_v^*}{m_v^* + m_c^*}(h\nu - E_g). \qquad (8.47)$$

Analogamente,

$$E_{V0} - E_1 = \frac{m_c^*}{m_v^* + m_c^*}(h\nu - E_g).$$

Así pues acabamos de relacionar el nivel E_2 (o E_1) con la frecuencia del fotón emitido o absorbido en la transición

Sabemos que la densidad de estados entre E_2 y E_2 +dE_2 viene dada por $g(E_2)dE_2$. Supongamos ahora que *todos* los estados en el intervalo dE_2 estuvieran ocupados por electrones, y que *todos* realizaran transiciones entre dE_2 y dE_1. ¿Cuál sería la densidad de fotones emitidos en el intervalo $d\nu$ correspondiente?. Es decir, queremos pasar de la densidad de estados a la densidad de frecuencias.

Para responder a esta cuestión, consideremos primero un caso hipotético en el que el semiconductor fuera bidimensional. Entonces, el vector de onda de los electrones tendría también sólo 2 dimensiones (2D) y la representación $E(\mathbf{k})$ la podemos "ver" en la Figura 8.10.2 . Los electrones con energías E_2 estarían sobre la circunferencia designada como E_2 y sus valores k_x,k_y determinan puntos que están sobre la

proyección de dicha circunferencia en el plano k_x,k_y, y que es la circunferencia interior dibujada. Análogamente pasa con E_2+dE_2 y las circunferencias interiores.

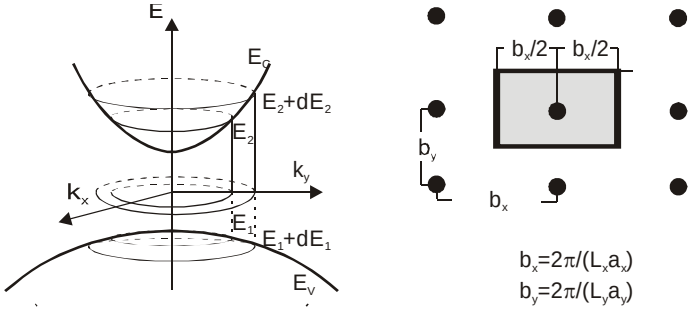


Figura 8.10.2 Representación en 2 dimensiones de las bandas de energía, y forma de obtener el área asociada a cada estado k

Todos los electrones con energías entre E_2 y E_2+dE_2 corresponderán a estados que estén entre ambas circunferencias. Teniendo en cuenta que se puede demostrar que en 2D el área que ocupa cada estado es $(2\pi)^2/S_C$, (siendo S_C $[= (N_x a_x)\cdot(N_y a_y)]$ la superficie del cristal, N_x, N_y el número de átomos del cristal en las direcciones x, y y a_x, a_y la distancia interatómica en dichas direcciones), el número de estados entre E_2 y E_2+dE_2 lo podemos obtener dividiendo el área entre las circunferencias en el plano k_x ,k_y por el área que ocupa cada estado.

Pues bien, esto mismo podemos hacer con tres dimensiones, reemplazando áreas por volúmenes. Obviamente, no podemos dibujarlo, por lo que hemos de recurrir a la abstracción del caso 2D. La superficie entre las dos circunferencias del caso 2D se transforma en el volumen entre dos esferas cuyas energías valen E_2 y E_2+dE_2. Este volumen lo podemos obtener fácilmente si tenemos en cuenta que a una esfera de energía E_2 le corresponde un radio k_2 y un volumen V_2 dados por:

$$E_2 - E_{C0} = \frac{\hbar^2 k^2}{2\,m_c^*} \Rightarrow k = \left(\frac{2\,m_c^*}{\hbar^2}\right)^{\frac{1}{2}} \left(E_2 - E_{C0}\right)^{1/2}; \qquad V_2 = \frac{4}{3}\pi\left(\frac{2\,m_c^*}{\hbar^2}\right)^{\frac{3}{2}} \left(E_2 - E_{C0}\right)^{3/2}. \tag{8.48}$$

Si ahora incrementamos E_2 hasta E_2+dE_2, el volumen V_2 experimenta un incremento dV_2 que se obtiene derivando (8.48):

$$dV_2 = 2\pi\left(\frac{2\,m_c^*}{\hbar^2}\right)^{\frac{3}{2}} \left(E_2 - E_{C0}\right)^{1/2}\, dE_2\,. \tag{8.49}$$

Como cada estado ocupa un volumen $(2\pi)^3/V$ en el espacio **k**, siendo V el volumen del cristal, el número de estados en dV_2 lo obtendremos dividiendo dV_2 por lo que ocupa cada estado:

$$dN = \frac{dV_2}{\frac{(2\pi)^3}{V}} = \frac{V}{4\pi^2}\left(\frac{2m_c^*}{\hbar^2}\right)^{\frac{3}{2}}\left(E_2 - E_{C0}\right)^{1/2}\, dE_2\,, \tag{8.50}$$

y teniendo en cuenta (8.47),

$$dN = \frac{V}{4\pi^2}\left(\frac{2m_c^*}{\hbar^2}\right)^{\frac{3}{2}}\left(\frac{m_v^*}{m_c^*+m_v^*}\right)^{\frac{1}{2}}\left(h\nu - E_g\right)^{1/2}\left(\frac{m_v^*}{m_c^*+m_v^*}\right) d(h\nu)\,, \tag{8.51}$$

A este número, referido a la unidad de volumen (es decir, dN/V) se le llama *densidad conjunta ("joint density") de estados*, $\rho_j(h\nu)$:

$$\rho_j(h\nu) = \frac{1}{4\pi^2}\left(\frac{2m_r^*}{\hbar^2}\right)^{\frac{3}{2}}\left(h\nu - E_g\right)^{1/2}\,; \quad m_r^* = \frac{m_c^* m_v^*}{m_c^*+m_v^*}\,. \tag{8.52}$$

$\rho_j(h\nu)\ d(h\nu)$ representa el número por unidad de volumen de transiciones con emisión de radiación con energía entre $h\nu$ y $h\nu+d(h\nu)$ que se producirían si los estados de partida estuvieran completamente llenos y los de llegada completamente vacíos. De esta manera ya hemos conseguido poner las energías en función de $h\nu$. Conviene remarcar aquí que la variable que interviene en ρ_j es $h\nu$, o sea, la energía. Si queremos pasarlo a frecuencia, más útil en nuestro caso, habremos de tener en cuenta que:

$$\rho_j(\nu)\ d\nu = h\,\rho_j(h\nu)\ d(h\nu)\,. \tag{8.53}$$

Ahora ya estamos en condiciones de calcular la ganancia de un semiconductor. Recuperemos las ecuaciones de Einstein (8.9)-(8.11) y planteemos un razonamiento análogo. Sea R_{12} el número de transiciones por unidad de tiempo y unidad de volumen que tienen lugar desde la banda de valencia a la de conducción, en el intervalo de frecuencias entre ν y $\nu+d\nu$. Estas transiciones dan cuenta de los procesos de absorción que tienen lugar en las condiciones citadas. De acuerdo con la fórmula de Einstein, R_{12} será igual al producto de:

1. Un coeficiente de proporcionalidad B_{12}
2. La densidad de energía $\rho(\nu)$ en el intervalo de frecuencias citado
3. El número de transiciones posibles en este intervalo de frecuencia si los estados de partida estuvieran completamente llenos y los de llegada completamente vacíos, $\rho_j(\nu)\,d\nu$
4. La probabilidad de que el estado de partida este realmente ocupado $f_v(E_1)$ por la probabilidad de que el estado de llegada esté vacío, $1\text{-}f_c(E_2)$:

$$R_{12} = B_{12} \cdot \rho(\nu) \cdot \rho_j(\nu)\ d\nu \cdot f_v(E_1) \cdot \left[1 - f_c(E_2)\right]. \tag{8.54}$$

De la misma forma, si R_{21} representan las transiciones producidas por emisión estimulada que tienen lugar por unidad de tiempo y volumen en el mismo intervalo de frecuencias, tendremos:

$$R_{21} = B_{21} \cdot \rho(\nu) \cdot \rho_j(\nu)\ d\nu \cdot f_c(E_2) \cdot \left[1 - f_v(E_1)\right]. \tag{8.55}$$

El número neto de transiciones con emisión de fotones es:

$$R_{21} - R_{12} = B_{21} \cdot \rho(\nu) \cdot \rho_j(\nu)\ d\nu \cdot \left[f_c(E_2) - f_v(E_1)\right]. \tag{8.56}$$

Aquí se ha tenido en cuenta que $B_{21} = B_{12}$ ya que la degeneración de los estados es la misma en ambos casos. Si este valor lo multiplicamos por $h\nu$ obtenemos la energía generada por unidad de tiempo y unidad de volumen, es decir la potencia neta generada por unidad de volumen.

Si ahora calculamos la potencia incidente por unidad de área, esto es $I(\nu)\, d\nu$, la intensidad en el intervalo de frecuencias $d\nu$, teniendo en cuenta la definición (8.22) de $\gamma(\nu)$ podremos obtener su valor

$$I(\nu)\ d\nu = \rho(\nu)\, d\nu \cdot c / n_g\,, \tag{8.57}$$

siendo c la velocidad de la luz en el vacío. Como

$$\gamma = \frac{\text{Incremento de potencia generada por unidad de volumen}}{\text{Intensidad incidente}}\,, \tag{8.58}$$

reemplazando tenemos:

$$\gamma(\nu) = \frac{h\nu\left(R_{21} - R_{12}\right)}{I(\nu)\ d\nu} = B_{21} \cdot h\nu \cdot \frac{n_g}{c} \cdot \rho_j(\nu) \cdot \left[f_c(E_2) - f_v(E_1)\right]. \tag{8.59}$$

Teniendo en cuenta la relación

$$B_{21} = A_{21}\, \frac{c^3}{8\pi h \nu^3 n^2 n_g}\,, \tag{8.60}$$

podemos poner

$$\gamma(\nu) = A_{21} \cdot \frac{c^2}{8\pi n^2 \nu^2} \cdot \rho_j(\nu) \cdot \left[f_c(E_2) - f_v(E_1)\right]. \tag{8.61}$$

Otra forma equivalente es:

$$\gamma(\nu) = A_{21} \cdot f_c(E_2) \cdot \left[1 - f_v(E_1)\right] \cdot \frac{c^2}{8\pi n^2 \nu^2} \cdot \rho_j(\nu) \cdot \left[1 - \frac{f_v(E_1) \cdot \left[1 - f_c(E_2)\right]}{f_c(E_2) \cdot \left[1 - f_v(E_1)\right]}\right] \tag{8.62}$$

Y operando un poco más:

$$\gamma(\nu) = A_{21} \cdot f_c(E_2) \cdot \left[1 - f_v(E_1)\right] \cdot \frac{c^2}{8\pi n^2 \nu^2} \cdot \rho_j(\nu) \cdot \left[1 - e^{\frac{h\nu - (E_{Fn} - E_{Fp})}{k_B T}}\right] . \quad (8.63)$$

Esta última expresión muestra un aspecto muy importante: Todos los factores son siempre positivos excepto el último corchete. Este puede ser negativo si $h\nu > E_{Fn} - E_{Fp}$. En este caso, en lugar de ganancia tendremos atenuación (absorción) de la radiación. Por lo tanto, para tener ganancia en un semiconductor intrínseco tendremos que seleccionar las concentraciones de portadores para que, a aquellas frecuencias que deseemos tener ganancia,

$$E_g < h\nu < E_{Fn} - E_{Fp} . \quad (8.64)$$

Este resultado lo habíamos deducido para T=0 K. Ahora se amplia a cualquier temperatura. Si estamos en equilibrio termodinámico, $E_{Fn} = E_{Fp}$, y por lo tanto, no puede haber ganancia. Para tenerla hemos de generar una situación de no equilibrio, por lo general bastante fuerte.

Si observamos la expresión (8.61) y el hecho de que $\rho_j(\nu)$ depende de ν de la forma $(h\nu - E_g)^{1/2}$, se suele representar $\gamma(\nu)$ por:

$$\gamma(\nu) = K(h\nu - E_g)^{1/2} \left[f_c(E_2) - f_v(E_1)\right]. \quad (8.65)$$

La constante K se determina experimentalmente, y E_2 y E_1 vienen dados por (8.47). A la vista de esta expresión se puede pensar que se ha eliminado la parte de dependencia con la frecuencia debida al término ν^2 en el denominador de (8.63), y así es. Ello se debe a que en el intervalo de frecuencias de interés, es decir, aquellas en las que puede haber ganancia, la influencia debida al citado término del denominador es relativamente poco importante.

8.11 Perfil de Emisión Espontánea

Además de relacionar el coeficiente de ganancia con el de absorción, podemos relacionar aquel con el espectro de emisión espontánea que resulta de la recombinación de los electrones y los huecos, y que se puede medir con bastante facilidad experimentalmente. El razonamiento para establecer tal relación es como sigue: Supongamos que tenemos un semiconductor contenido en una cavidad cerrada que actúa como cuerpo negro, ambos en equilibrio termodinámico a una temperatura T. Dentro de la cavidad, la densidad de energía de la radiación viene dada por la fórmula de Plank (8.12) y que repetimos por más comodidad a continuación:

$$\rho(\nu)\, d\nu = \frac{8\pi\nu^2 n^2 n_g}{c^3} \cdot \frac{h\nu\, d\nu}{\exp(h\nu / k_B T) - 1} . \quad (8.66)$$

Esta energía, al incidir sobre el semiconductor y ser absorbida por él, producirá transiciones de electrones desde la banda de valencia a la de conducción, que serán reemitidos a continuación de nuevo a la banda de valencia por recombinación ya que en equilibrio termodinámico, la concentración de huecos y electrones es constante. Además, en esta situación, la energía absorbida por el semiconductor del campo de radiación habrá de ser igual a la que devuelve a partir de la recombinación de electrones y huecos (*principio del equilibrio microscópico*).

Sea $R(\nu)d\nu$ el número de fotones generados por unidad de tiempo y de volumen, por los procesos espontáneos de recombinación, con frecuencias entre ν y $\nu+d\nu$. Si multiplicamos esta cantidad por $h\nu$ obtenemos la energía emitida por unidad de tiempo y volumen (=potencia emitida por unidad de volumen) en el intervalo de frecuencias. Por lo dicho anteriormente, esta potencia ha de ser igual a la absorbida, y atendiendo a (8.22),

$$h\nu\, R(\nu)d\nu = \left[\alpha(\nu)\right]_{eq} \cdot I(\nu)d\nu = \alpha(\nu)\cdot(c/n_g)\cdot\rho(\nu)d\nu\,, \tag{8.67}$$

en donde hemos puesto el subíndice "eq" al parámetro α para subrayar la condición de equilibrio termodinámico. Reemplazando (8.66) y operando:

$$\left[\alpha(\nu)\right]_{eq} = R(\nu)\cdot\frac{c^2}{8\pi\nu^2 n^2}\cdot\left(e^{\frac{h\nu}{k_B T}}-1\right). \tag{8.68}$$

La condición de equilibrio termodinámico implica que $E_{Fn} = E_{Fp} = E_F$. Así pues, podemos poner:

$$\left[\alpha(\nu)\right]_{eq} = \left[-\gamma(\nu)\right]_{E_{Fn}=E_{Fp}} =$$

$$= A_{21}\cdot f(E_2)\cdot\left[1-f(E_1)\right]\cdot\frac{c^2}{8\pi\nu^2 n^2}\cdot\rho_j(\nu)\cdot\left(e^{\frac{h\nu}{k_B T}}-1\right) \tag{8.69}$$

Lo que nos permite identificar

$$R(\nu) = A_{21}\cdot f(E_2)\cdot\left[1-f(E_1)\right]\cdot\rho_j(\nu)\,, \tag{8.70}$$

y relacionar el coeficiente de ganancia con datos experimentales:

$$\gamma(\nu) = R(\nu)\cdot\frac{f_c(E_2)\cdot\left[1-f_v(E_1)\right]}{f(E_2)\cdot\left[1-f(E_1)\right]}\cdot\frac{c^2}{8\pi n^2\nu^2}\cdot\left[1-e^{\frac{h\nu-(E_{Fn}-E_{Fp})}{k_B T}}\right]. \tag{8.71}$$

Es bastante habitual definir, de forma semiempírica,

$$R(\nu)\cdot\frac{f_c(E_2)\cdot\left[1-f_v(E_1)\right]}{f(E_2)\cdot\left[1-f(E_1)\right]} = \frac{n}{\tau_r}g(\nu)\,, \tag{8.72}$$

en donde n es la densidad de electrones en la banda de conducción, τ_r es la vida media de recombinación espontánea y $g(\nu)$ una función que

representa la distribución espectral de la emisión espontánea y el cociente entre las funciones de ocupación en el equilibrio y fuera de él. Así:

$$\gamma(\nu) = \frac{n}{\tau_r} \cdot \frac{c^2}{8\pi n^2 \nu^2} \cdot g(\nu) \cdot \left[1 - e^{\frac{h\nu - (E_{Fn} - E_{Fp})}{k_B T}} \right]. \qquad (8.73)$$

8.12 Ejemplo: Obtención de una condición umbral para obtener ganancia en un semiconductor

Vamos a calcular, a modo de ejemplo, una condición mínima para que en un semiconductor tal como arseniuro de galio (GaAs) el coeficiente de ganancia sea mayor que cero en algún intervalo de frecuencias. Para ello tendremos que conseguir que los pseudoniveles de Fermi se encuentren a una distancia en energía mayor que el gap del GaAs (E_g = 1,43 eV). La forma física de conseguirlo es hacer que haya muchos electrones en la banda de conducción y muchos huecos en la de valencia a la vez, lo que obviamente, está muy alejado de la condición de equilibrio termodinámico. Esta situación se denomina *inversión de población*, ya que es la inversa de las "normales" que habitualmente tenemos.

Para obtener la densidad de electrones en la banda de conducción tenemos que multiplicar la densidad de estados entre E_2 y E_2+dE_2, $g_c(E_2)dE_2$, por la probabilidad de que estos estados estén ocupados por electrones $f_c(E_2)$ y sumar para todo el intervalo de energía que incluye la banda de conducción, que va desde E_{C0} hasta un valor muy grande E_{CM} que, por lo general, no conocemos y que podemos tomar como infinito sin error apreciable, ya que para él la probabilidad de ocupación será prácticamente nula:

$$n = \int_{E_{C0}}^{\infty} f_c(E_2)\, g_c(E_2) dE_2 \,. \qquad (8.74)$$

Análogamente, la densidad de huecos en la banda de valencia la podemos expresar como:

$$p = \int_{-\infty}^{E_{V0}} \left[1 - f_v(E_1)\right] \cdot g_v(E_1) dE_1 \,. \qquad (8.75)$$

Aquí es el valor mínimo de la banda de valencia el que se toma muy alejado, a menos infinito.

La densidad de estados en la banda de conducción no es sino el número de estados por unidad de volumen que hay entre E_2 y E_2+dE_2, y la podemos obtener directamente de (8.50) teniendo en cuenta que $g_c(E_2)dE_2 = 2\ dN/V$ ya que hay dos electrones por cada estado debido a la degeneración por el spin. Por lo tanto,

$$n = \frac{1}{2\pi^2}\left(\frac{2m_c^*}{\hbar^2}\right)^{3/2} \int_{E_{C0}}^{\infty} \frac{(E_2 - E_{C0})^{1/2}}{1 + e^{\frac{E_2 - E_{Fn}}{k_B T}}} dE_2 \,. \tag{8.76}$$

Siguiendo un proceso completamente similar al utilizado para la obtención de $g_c(E_2)$ se puede obtener $g_v(E_1)$, y la expresión que da la densidad de huecos en la banda de valencia es:

$$p = \frac{1}{2\pi^2}\left(\frac{2m_v^*}{\hbar^2}\right)^{3/2} \int_{-\infty}^{E_{V0}} \frac{(E_{V0} - E_1)^{1/2}}{1 + e^{-\frac{E_1 - E_{Fp}}{k_B T}}} dE_1 \,. \tag{8.77}$$

El signo menos en el exponente del denominador de la expresión anterior aparece porque para obtener la densidad de huecos hay que multiplicar por la probabilidad de que *no* haya electrones en los estados considerados en la banda de valencia, $[1\text{-}f_v(E_1)]$. Ambas integrales se pueden tratar de forma muy similar. En efecto, si llamamos

$$\begin{aligned} \frac{E_2 - E_{C0}}{k_B T} &= x \qquad \frac{E_{V0} - E_1}{k_B T} = y \\ \frac{E_{Fn} - E_{C0}}{k_B T} &= \zeta \qquad \frac{E_{V0} - E_{Fp}}{k_B T} = \theta \end{aligned} \tag{8.78}$$

y operamos las expresiones anteriores de *n* y *p* tenemos:

$$\begin{aligned} n &= \frac{1}{2\pi^2}\left(\frac{2m_c^* k_B T}{\hbar^2}\right)^{3/2} \int_0^{\infty} \frac{x^{1/2}}{1 + e^{x-\zeta}} dx \\ p &= \frac{1}{2\pi^2}\left(\frac{2m_v^* k_B T}{\hbar^2}\right)^{3/2} \int_0^{\infty} \frac{y^{1/2}}{1 + e^{y-\theta}} dy \end{aligned} \,. \tag{8.79}$$

Las integrales son las mismas en las dos expresiones, excepción hecha de ζ y θ, ya que *x* e *y* son variables de integración que, por si mismas, no significan nada. Este tipo de integrales se denominan *integrales de Fermi de orden 1/2*, y forman parte de un conjunto más amplio de integrales de Fermi de orden *n/2* definidas por:

$$\Im_{\frac{n}{2}}(\xi) = \int_0^{\infty} \frac{z^{n/2}}{1 + e^{z-\xi}} dz \,. \tag{8.80}$$

Para estas integrales no se conoce ninguna expresión analítica exacta, para ningún valor de $n \neq 0$ de modo que su cálculo se realiza bien mediante una aproximación analítica del integrando suficientemente precisa 0, o integrándola numéricamente para distintos valores de ξ. Nosotros utilizaremos aquí el segundo método, aunque el primero es más general.

Para mantener una conexión más próxima con la electrónica de los dispositivos vista en capítulos anteriores, vamos a dejar el tratamiento

de las integrales de Fermi y multipliquemos y dividamos las expresiones de *n* y *p* por *exp(ζ)* y *exp(θ)* respectivamente, y pongámoslo de la forma:

$$\begin{aligned} n &= \frac{1}{2\pi^2}\left(\frac{2m_c^* k_B T}{\hbar^2}\right)^{3/2} e^{-\frac{E_{C0}-E_{Fn}}{k_B T}} \int_0^\infty \frac{x^{1/2}}{e^{\zeta}+e^{x}}dx \\ p &= \frac{1}{2\pi^2}\left(\frac{2m_v^* k_B T}{\hbar^2}\right)^{3/2} e^{\frac{E_{V0}-E_{Fp}}{k_B T}} \int_0^\infty \frac{y^{1/2}}{e^{\theta}+e^{y}}dy \end{aligned} \quad . \tag{8.81}$$

En régimen de funcionamiento "normal" de la mayoría de los dispositivos electrónicos habituales, la situación es tal que $E_{Fn} - E_{C0} < 0$ y $E_{V0} - E_{Fp} < 0$, por lo que siempre, en esos casos, $exp(\zeta) < exp(x)$ y $exp(\theta) < exp(y)$, y, también casi siempre, el símbolo "menor" se puede reemplazar por "mucho menor". Así,

$$\int_0^\infty \frac{x^{1/2}}{e^{\zeta}+e^{x}}dx \cong \int_0^\infty \frac{x^{1/2}}{e^{x}}dx = \frac{\sqrt{\pi}}{2}, \tag{8.82}$$

y lo mismo se obtiene para huecos. Pongamos (8.81) como:

$$\begin{aligned} n &= \frac{1}{2\pi^2}\cdot\left(\frac{2m_c^* k_B T}{\hbar^2}\right)^{3/2}\cdot\frac{\sqrt{\pi}}{2}\cdot e^{-\frac{E_{C0}-E_{Fn}}{k_B T}}\cdot\left(\frac{2}{\sqrt{\pi}}\int_0^\infty \frac{x^{1/2}}{e^{\zeta}+e^{x}}dx\right) \\ p &= \frac{1}{2\pi^2}\cdot\left(\frac{2m_v^* k_B T}{\hbar^2}\right)^{3/2}\cdot\frac{\sqrt{\pi}}{2}\cdot e^{\frac{E_{V0}-E_{Fp}}{k_B T}}\cdot\left(\frac{2}{\sqrt{\pi}}\int_0^\infty \frac{y^{1/2}}{e^{\theta}+e^{y}}dy\right) \end{aligned} . \tag{8.83}$$

Los productos de constantes previos a los términos exponenciales representan lo que normalmente se denomina N_C y N_V, *densidad equivalente de estados en la banda de conducción y de valencia* respectivamente. Así pues:

$$\begin{aligned} n &= N_C\cdot e^{-\frac{E_{C0}-E_{Fn}}{k_B T}}\cdot\left(\frac{2}{\sqrt{\pi}}\int_0^\infty \frac{x^{1/2}}{e^{\frac{E_{C0}-E_{Fn}}{k_B T}}+e^{x}}dx\right) \\ p &= N_V\cdot e^{\frac{E_{V0}-E_{Fp}}{k_B T}}\cdot\left(\frac{2}{\sqrt{\pi}}\int_0^\infty \frac{y^{1/2}}{e^{\frac{E_{V0}-E_{Fp}}{k_B T}}+e^{y}}dy\right) \end{aligned} . \tag{8.84}$$

Si consideramos una temperatura de 27 °C (T = 300 K; k_BT = 0.0259 eV); $m_c^* = 0.067\, m_0$ y $m_v^* = 0.55\, m_0$ (m_0 es la masa en reposo del electrón. Estos datos son propios del GaAs), obtenemos $N_C = 4.36\ 10^{17}\ cm^{-3}$, $N_V = 1.02\ 10^{19}\ cm^{-3}$. Para seguir con los cálculos, llamaremos I al termino entre paréntesis en ambos casos, y utilizaremos la tabla Tabla 8.7.1, de la que podemos deducir algunas consecuencias importantes:

$exp[(E_{C0}-E_{Fn})/k_BT]$ o $exp[(E_{Fp}-E_{V0})/k_BT]$	$(E_{C0}-E_{Fn})/k_BT$ o $(E_{Fp}-E_{V0})/k_BT$	I	n (cm^{-3})	p (cm^{-3})
10^3	6.91	1.000	$4.36\cdot10^{14}$	$1.02\cdot10^{16}$
10^2	4.61	0.996	$4.34\cdot10^{15}$	$1.02\cdot10^{17}$
50	3.91	0.993	$8.66\cdot10^{15}$	$2.04\cdot10^{17}$
20	3.00	0.983	$2.14\cdot10^{16}$	$5.04\cdot10^{17}$
10	2.30	0.967	$4.22\cdot10^{16}$	$9.92\cdot10^{17}$
7.75	2.05	0.957	$5.40\cdot10^{16}$	$1.27\cdot10^{18}$
5	1.61	0.936	$8.16\cdot10^{16}$	$1.92\cdot10^{18}$
2	0.69	0.860	$1.87\cdot10^{17}$	$4.41\cdot10^{18}$
1	0.00	0.765	$3.34\cdot10^{17}$	$7.84\cdot10^{18}$
0.5	-0.69	0.641	$5.59\cdot10^{17}$	$1.31\cdot10^{19}$
0.2	-1.61	0.457	$9.96\cdot10^{17}$	$2.34\cdot10^{19}$
0.129	-2.05	0.373	$1.27\cdot10^{18}$	$2.99\cdot10^{19}$
0.1	-2.30	0.329	$1.43\cdot10^{18}$	$3.37\cdot10^{19}$

Tabla 8.12.1 Cálculo de los valores de *n* y *p* para distintas posiciones de los pseudoniveles de Fermi. Las filas sombreadas indican posiciones de los pseudoniveles de Fermi dentro de las bandas prohibidas

- Para situar los pseudoniveles de Fermi en una determinada posición respecto de los extremos de las bandas hacen falta muchos más huecos que electrones. Por lo tanto, sería conveniente inyectar electrones en un material muy dopado de tipo *p*. Sin embargo, en los materiales dopados a los niveles altos de la tabla (*~10^{19} cm^{-3}*), los niveles de energía creados por las impurezas ya no son "individuales", sino que se mezclan entre ellos creando "bandas de niveles", y también interaccionan con las bandas de energía cambiando la forma del gap, de manera que éste no queda claramente definido entre estados permitidos y no permitidos, lo que conlleva que la regla de conservación de **k** ya no tiene que cumplirse necesariamente. Para evitar estas cuestiones, vamos a seguir con nuestro modelo sencillo.
- Si el dopado no es muy alto, entonces la conservación de la neutralidad eléctrica obligará a que $n \approx p$ y estos provendrán de la inyección eléctrica. Por lo tanto, para conseguir situar los niveles de Fermi de manera que $E_g < E_{Fn} - E_{Fp}$ necesitaremos un valor *mínimo* de $n = p = 1.27\ 10^{18}\ cm^{-3}$. Este valor se deduce de

los de la tabla, y es el menor que cumple que $n = p$ y $E_{Fn} - E_{Fp} = E_g$. Esta condición *mínima* para poder tener ganancia se denomina *condición umbral* para el funcionamiento del láser, y es uno de los parámetros más importantes en cualquier cálculo sobre el funcionamiento de estos dispositivos. Lógicamente, si aumentáramos los valores de *n* y *p* obtendríamos ganancia neta

Por lo tanto, hemos de conseguir esas concentraciones de electrones y huecos. Estas densidades tan altas darán lugar a una recombinación muy intensa que emitirá luz -eso es lo que queremos-, y que provocará una disminución de la densidad de electrones que, al ser banda a banda, podemos tomar proporcional al producto de *n* y *p*. Si al mismo tiempo aportamos una densidad de electrones por unidad de tiempo dada por *G (bombeo de portadores)*, la densidad de electrones en la banda de conducción (y huecos en la de valencia) variará como:

$$\frac{dn}{dt} = -R \cdot n \cdot p + G\,, \tag{8.85}$$

en donde *R* es el coeficiente de proporcionalidad de la recombinación, y para estas densidades de huecos y electrones en GaAs se puede tomar como $R = 2{\cdot}10^{-10}\ cm^{-3}/sec$. En régimen estacionario, para mantener estas concentraciones necesitaremos un aporte (*bombeo*) *G* dado por:

$$\frac{dn}{dt} = 0 \Rightarrow G = R \cdot n \cdot p = 2 \cdot 10^{-10} \cdot \left(1.27 \cdot 10^{18}\right)^2 = 3.23 \cdot 10^{26} cm^{-3} s^{-1}\,. \tag{8.86}$$

Como la densidad de electrones inyectados es igual a la de huecos, *G* se expresa normalmente como:

$$G = 3.23 \cdot 10^{26}\, \frac{pares \quad e-h}{cm^3\, s}\,. \tag{8.87}$$

Si admitimos ahora que la recombinación se produce en una distancia *d* típica de *1 μm*, (más adelante veremos como obtener esta distancia) la densidad de corriente *J* necesaria para tener este valor de *G* es:

$$J = q \cdot d \cdot G = 1.6 \cdot 10^{-19} \cdot C \cdot 10^{-4} \cdot cm \cdot 3.23 \cdot 10^{26}\, \frac{1}{cm^3\, s} = 5.17\, \frac{kA}{cm^2}\,. \tag{8.88}$$

(q es el valor absoluto de la carga del electrón).

Este valor de la corriente es considerablemente grande, y de todas las magnitudes que intervienen, la que parece más fácil de manipular sería la distancia d en la que hay recombinación, ya que el resto son parámetros propios del semiconductor. Así se hace, y en los apartados siguientes veremos cómo.

De las expresiones de *n* y *p* dadas por (8.83), podemos deducir que estos parámetros tienen una dependencia muy grande con la temperatura, y, por ello, también lo será la densidad de corriente umbral.

Especialmente importante es el término de la forma $exp(-\Delta E/k_B T)$, y por lo general se admite de forma semiempírica que la densidad de corriente umbral a una temperatura T, $J(T)$, se relaciona con la correspondiente a otra temperatura T_0, J_0, por

$$J(T) = J_0 \cdot e^{\frac{T-T_0}{T_0}}, \tag{8.89}$$

en donde T_0 suele tener un valor aproximado de *400 K* para un buen láser.

8.13 Diodos láser de homounión

Vamos a ver ahora como se aplicaría lo dicho anteriormente a un diodo láser, y para empezar, consideraremos un diodo exclusivamente de GaAs, es decir un dispositivo formado por un único semiconductor, esto es, un láser de homounión.

La Figura 8.13.1 muestra una estructura típica de uno de tales láseres. Los que así se construyeron era básicamente una unión P-N (de GaAs, aunque desde el punto de vista del tratamiento teórico el semiconductor puede ser otro). Lo importante es que toda la estructura es del mismo semiconductor.

La zona sombreada representa la zona de carga espacial, y la rayada, la zona en la que puede haber emisión de luz, no necesariamente con coeficiente de ganancia positivo. La longitud de esta zona viene determinada por la dispersión de la corriente debida, por ejemplo, a la desigual superficie de los contactos metálicos. La anchura la podemos determinar a partir del comportamiento general de una unión p-n.

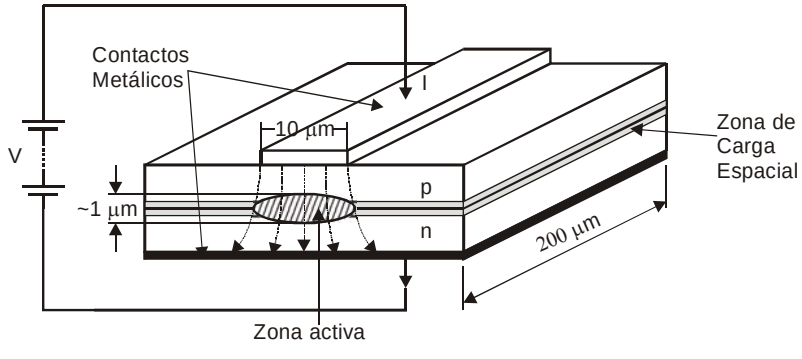


Figura 8.13.1 Esquema de un diodo láser de GaAs con dimensiones típicas. La zona activa tiene una anchura aproximadamente igual al contacto metálico superior, y es mayor que la zona de carga espacial

Debido a la alta inyección de portadores que hace falta para situar los pseudoniveles de Fermi dentro de las bandas de conducción y de

valencia, los dos materiales están degenerados. Para simplificar el problema, consideraremos que aún así vale la representación típica de los semiconductores con niveles de energía creados por las impurezas bien definidos (no bandas de impurezas), con los extremos de las bandas de energía perfectamente delimitados y que se cumple la regla de conservación de **k**.

Otro efecto de la alta inyección sobre las uniones p-n (en general) es que la zona de carga espacial prácticamente desaparece, y en cambio, los electrones inyectados en la zona p penetran profundamente en ella, y ahí tiene lugar la recombinación con los huecos, mayoritarios en el material tipo p. Lo mismo sucede con los huecos en la zona n. Las profundidades de esta penetración podemos estimarlas del orden de las longitudes de difusión $L_n = (D_n\ \tau_n)^{1/2}$ y $L_p = (D_p\ \tau_p)^{1/2}$, siendo D_n *y* D_p los coeficientes de difusión de huecos y de electrones respectivamente, y τ_n y τ_p sus vidas medias. Para el GaAs , a estas concentraciones, podemos tomar como valores típicos:

$\mu_n = 600\ \text{cm}^2/(\text{V}\cdot\text{s}) \Rightarrow D_n = \mu_n \cdot (k_B T/q) = 15.51\ \text{cm}^2/\text{s}$

$\mu_p = 30\ \text{cm}^2/(\text{V}\cdot\text{s}) \Rightarrow D_p = \mu_p \cdot (k_B T/q) = 0.78\ \text{cm}^2/\text{s},$

con lo que vemos que, si admitimos que la vida media de los electrones y de los huecos, tendremos que:

$$L_n = \sqrt{D_n \tau_n} \approx \sqrt{20 \cdot D_p \tau_p} = 4.47 \cdot L_p\,, \tag{8.90}$$

es decir, la longitud de difusión de los electrones será unas 4.5 veces la de los huecos. La Figura 8.13.2 muestra la forma de dichas concentraciones.

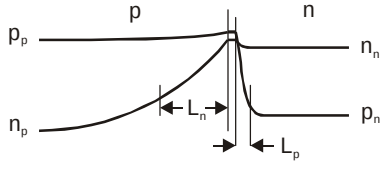


Figura 8.13.2 Representación esquemática de las concentraciones de huecos y de electrones en las zonas p y n en condiciones de alta inyección. La longitud de difusión de los electrones en la zona p es apreciablemente mayor que la de los huecos en la n.

Para hacer una estimación aproximada de los valores numéricos de las magnitudes características de este tipo de dispositivos podemos despreciar la longitud de difusión de huecos y admitir que toda la recombinación tiene lugar en una distancia *d* tal que

$$d \approx L_n\,, \tag{8.91}$$

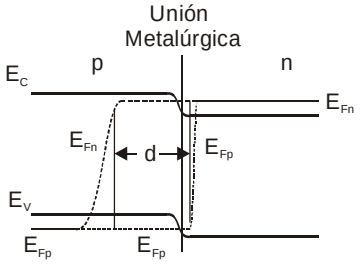


Figura 8.13.3 Diagrama de la forma de los pseudoniveles de Fermi en una unión p-n de GaAs. La variación espacial de los mismos se produce aproximadamente en el entorno de la longitud de difusión

Para calcular los parámetros propios de un diodo hemos de conocer las concentraciones de portadores, o en su lugar, las posiciones energéticas de los pseudoniveles de Fermi en cada punto de la estructura. Estos datos no son nada simples de obtener, así que para fijar ideas tomaremos para ellos unos valores razonables.

Los electrones en la zona p, n_p se recombinan con los huecos allí presentes en densidad p_p de acuerdo con

$$\frac{dn_p}{dt} = -\frac{n_p}{\tau_p} = -\left(\beta \cdot p_p\right) \cdot n_p \,. \tag{8.92}$$

Considerando los mismos valores que en el apartado anterior, esto es, que $p_p = 9 \cdot 10^{18}$ cm^{-3} y $\beta = 2 \cdot 10^{-10}$ cm^3/s obtenemos:

$$\tau_n = \frac{1}{1.8 \cdot 10^9} s = 0.56\, ns \Rightarrow L_n = \sqrt{15.51 \cdot 0.56 \cdot 10^{-9}} = 0.93 \mu m \,. \tag{8.93}$$

Si ahora admitimos que el pseudonivel de Fermi de electrones está 50 meV por encima del extremo inferior de la banda de conducción, y la temperatura de funcionamiento es de 350 K, entonces $(E_C - E_{Fn})/k_BT = -1.66$. De la Tabla 8.12.1, obtenemos $n \approx 9.96 \cdot 10^{17}$ cm^{-3}. Así pues, tenemos una densidad de electrones n que se recombina en una distancia $d \approx L_n$ a una velocidad n/τ_r La densidad de corriente J para mantener esta situación de no equilibrio es:

$$J = \frac{qnd}{\tau_r} = 18.49 \frac{\text{kA}}{\text{cm}^2} \,. \tag{8.94}$$

Este resultado es parecido al del apartado anterior, aunque algo diferente por los valores numéricos utilizados, y es típico de los láseres de homounión. Como se ve, J sigue siendo bastante grande, y para unas dimensiones como las dadas en la Figura 8.13.1, la intensidad de corriente que debería circular sería I=370 mA

De los razonamientos utilizados podemos deducir algunas consideraciones importantes:

- La distancia *d* en la que hay recombinación es un parámetro no controlable tecnológicamente, esto es, en el momento de diseñar el dispositivo. De nuevo vemos que si se consiguiera actuar sobre *d* se podría controlar *J*, y cualquier reducción de *d* repercutiría en una reducción de *J*, y por lo tanto, de *I*
- La distancia d es del orden de la longitud de onda de la luz amplificada. El haz de luz puede extenderse a una distancia mayor que *d*, de forma que no toda la intensidad de luz producida participará en la amplificación.
- La forma tan poco simétrica de la zona donde se produce luz hará que el haz emergente sea muy divergente (se "abra mucho" a medida que nos alejemos del dispositivo emisor).

Los dos primeros defectos se pueden corregir con el uso de heterouniones, propias de los láseres semiconductores actuales. El tercero de ellos se aplica a todos los láseres semiconductores, y es posiblemente la mayor desventaja de éstos frente a los láseres de gas o de estado sólido tipo YAG por ejemplo.

RESUMEN

En este capítulo se ha iniciado el estudio de los dispositivos optoelectrónicos, empezando con los conceptos básicos de la interacción luz-materia. Partiendo de las ecuaciones de Einstein se han definido los coeficientes de absorción y de ganancia de la intensidad luminosa en un material cualquiera, se ha revisado sucintamente la teoría de semiconductores para establecer las reglas que rigen las transiciones entre los estados de diferentes bandas, obteniendo la relación básica de conservación del momento cristalino, que distingue las transiciones permitidas de las que no lo están. De esta forma podemos explicar cómo se produce la amplificación de la luz en un semiconductor, es decir, como funciona un amplificador óptico. Si ahora esta luz amplificada la hacemos ir y volver entre dos espejos situados al inicio y final de la zona de amplificación de forma que se constituyan ondas estacionarias y podamos por lo tanto seleccionar su frecuencia tenemos un oscilador con luz: un láser. Como en todos los osciladores, se puede establecer una condición umbral para el arranque de la oscilación, y otra de mantenimiento de la misma. Hemos establecido tales condiciones, las hemos calculado en el caso de semiconductores de bandas esféricas (tipo GaAs), definiendo a la vez conceptos de uso más general como la densidad conjunta de estados, y se han determinado, la forma de la dependencia de la ganancia con la frecuencia de la radiación, así como el perfil de emisión espontánea. Finalmente, se ha aplicado todo este conocimiento para la determinación de la condición umbral de diodo láser de homounión de GaAs.

CUESTIONES Y PROBLEMAS

1. Un modelo algo más realista del GaAs puede obtenerse admitiendo que la banda de valencia está compuesta por dos bandas, denominadas banda de huecos ligeros y banda de huecos pesados, y se describen por la expresión (8.44), reemplazando m_v^* por m_{lh}^* y m_{hh}^* respectivamente (*lh = light holes*, *hh = heavy holes*). Estas bandas tienen la misma energía en $\mathbf{k}$=0 y se pueden tratar como bandas independientes. Pues bien, considérese una muestra de GaAs intrínseco con $m_e^* = 0.067\, m_0$, $m_{lh}^* = 0.067 m_0$, $m_{hh}^* = 0.55 m_0$ siendo m_0 la masa en reposo del electrón ($m_0 = 9.1 \cdot 10^{-31}$ Kg), y $E_g = 1.43$ eV. Mediante bombeo óptico se generan $5 \cdot 10^{18}$ cm^{-3} electrones en la banda de conducción (e igual densidad de huecos en total en las de valencia). Considerando T = 0 K

a) Determinar la posición del pseudonivel de Fermi de electrones E_{Fn} con respecto al extremo inferior de la banda de conducción (Rta: 0.1596 eV)

b) ¿Cual es la posición de E_{Fp} relativa a E_{V0}? (Rta: 0.0189 eV)

c) Obtener las densidades de huecos ligeros y pesados (Rta: $p_{lh} = 2 \cdot 10^{17} cm^{-3}$, $p_{ph} = 4.8 \cdot 10^{18} cm^{-3}$)

d) Aceptando que la energía de los electrones en la banda de valencia por encima de su mínimo de energía es toda ella energía cinética, ¿Cuál es la "velocidad" de los electrones en el estado de mayor energía ocupado?. (Rta: $9.14 \cdot 10^{17} cm/s$)

e) Repetir los apartados anteriores pero considerando que la temperatura es de T = 300 K. Observar las diferencias. (Para este apartado es conveniente utilizar alguna herramienta informática de cálculo numérico)

2. En una muestra de GaAs intrínseco a T = 0 K se generan pares electrón hueco por absorción de la radiación producida por un haz de luz de un láser de argón con λ =514.5 nm. Considerar que la potencia de bombeo es de 10^3 W/cm^3, que cada fotón absorbido crea un par electrón-hueco, que estos se recombinan de acuerdo con (8.85), con $\beta = 2 \cdot 10^{-10}$ cm^3/s, y que se ha alcanzado un régimen estacionario.

   a) ¿Cuál es la densidad de huecos y de electrones en sus respectivas bandas? (Rta: $3.6 \cdot 10^{15} cm^{-3}$)

   b) ¿Cuál es la distancia $E_{Fn} - E_{C0}$ (en eV)? (Rta: $1.28 \cdot 10^3$ eV)

   c) Repetir los apartados anteriores a T 300 K y comparar.

   (Tomar los datos necesarios del problema 1. Para el apartado c) es muy conveniente usar algún programa de cálculo numérico)

3. Un semiconductor intrínseco de GaAs se irradia con una onda tal que $h\nu - Eg = 0.05$ eV Suponiendo T = 0 K e ignorando los huecos ligeros,

   a) Identificar los niveles de energía en la banda de conducción y de valencia que pueden participar en la transición, admitiendo conservación de **k**. (Rta: $E_2 - E_{C0} = 0.044$ eV; EV0 - E1 = 0.0054 eV)

   b) Determinar la densidad mínima de pares electrón-hueco necesarios para conseguir ganancia a esa frecuencia. (Rta: $7.4 \cdot 10^{17}$ cm^{-3})

   c) Complicar el problema repitiendo los apartados anteriores sin despreciar la banda de huecos ligeros y a T = 300 K.

   (Tomar los datos necesarios del problema 1. Para el apartado c) es muy conveniente usar algún programa de cálculo numérico)

4. Las dimensiones de un semiconductor pueden llegar a ser tan pequeñas que la difusión de los portadores del punto en que se producen puede convertirse en un problema. La ecuación

que rige la generación y perdida de los poretadores puede ponerse como:

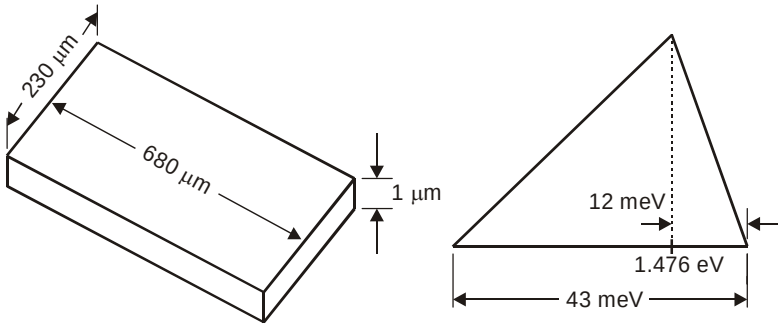
$$\frac{dn}{dt} = G - \beta \cdot n^2 - \frac{n}{\tau_D},$$

siendo τ_D la vida media de difusión, que para este problema tomaremos de 2.6 ns, y β el coefienciente de recombinación de huecos y electrones, cuyo valor será de $2 \cdot 10^{-10}$ cm^3/s.

a) ¿Cuál debe ser la concentración mínima de portadores para que la velocidad de recombinación exceda a la de difusión? (Rta: $1.9 \cdot 10^{18}$ cm^{-3})

b) Si se desea mantener una densidad de portadores de $3 \cdot 10^{18}$ cm^{-3} bombeando con un láser de argón que emite con λ =514,5 nm, y cada fotón crea un par electrón hueco, obtener la potencia absorbida por unidad de volumen de la radiación de bombeo. (Rta: $1.14 \cdot 10^9$ W/cm^3)

5. En la figura adjunta se muestra un esquema del perfil de emisión espontánea de un láser de GaAs. La longitud del mismo es de 680 μm, su índice de refracción es 3.6, la reflectividad de las aras es de 0.3, el coeficiente de absorción residual en el cristal es 10 cm^{-1}, y la vida media de recombinación, 1 ns.



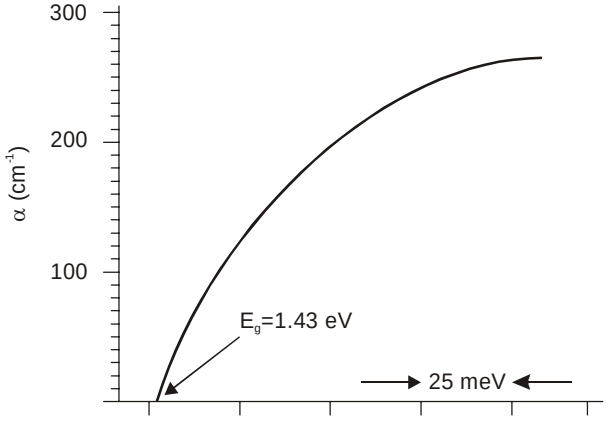
a) ¿Cuál es la longitud de onda para la máxima ganancia? (Rta: 0.84 μm)

b) ¿Cuál es la anchura a mitad de altura de altura (FWHM *= Full Width at Half Maximum*) del coeficiente de ganancia en Hz y cm^{-1}? (Rta: $5.1 \cdot 10^{12}$ Hz y 169 cm^{-1})

c) Determinar la densidad de portadores necesaria para llevar el láser a la condición umbral. (Rta: $6.5 \cdot 10^{15}$ cm^{-3})

d) Esta concentración de portadores debe mantenerse mediante algún tipo de bombeo -inyección de portadores, bombeo óptico u otros. Estimar la mínima potencia de bombeo para mantener una inversión de 10^{16} cm^{-3} en toda la estructura. (Rta: 369 mW)

6. La gráfica de la figura muestra el coeficiente de absorción a T =0 k de un semiconductor. Si la nuestra se somete a un bombeo óptico tal que $E_{Fn} - E_{C0} = 0.050$ eV y $E_{V0} - E_{Fp} = 2$ meV, encontrar el valor máximo del coeficiente de ganancia así como la energía de los fotones a la que se produce.



REFERENCIAS

J. T. Verdeyen, *Laser Electronics, 3rd Edition*. Prentice Hall, 1995

A. Yariv, *Quantum Electronics, 3rd Edition*. John Wiley and Sons, 1989

S.L. Chuang, *Physics of Optoelectronic Devices*. John Wiley and Sons, 1995

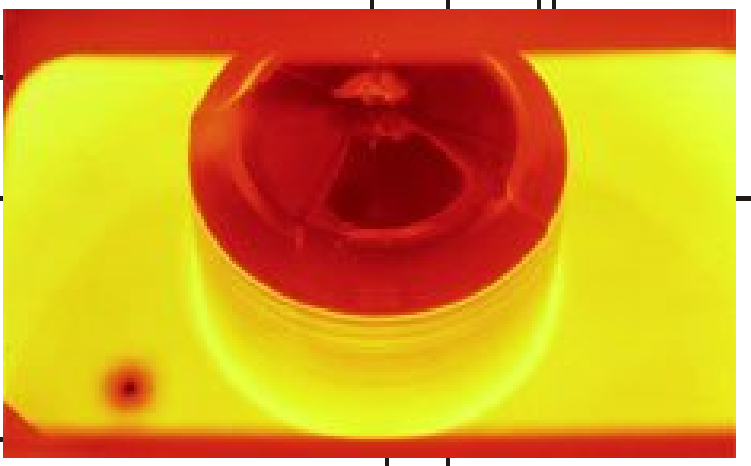
C. Kittel, *Introducción a la Física del Estado Sólido, 3ª Edición*. Ed. Reverté, S.A. Barcelona, 1975.

N. Ashcroft, M. Mermin. *Solid State Physics.* Holt, Rinehart and Winston, 1976

I. Melchor Ferrer. Tesis Doctoral. Departamento de Electrónica. Universidad de Granada, 1997

Capítulo 9

Crecimiento de Semiconductores.



Fabricación de una oblea de Silicio

Índice

Objetivos

- Describir los principales métodos de crecimiento en volumen de un cristal semiconductor.
- Explicar como se fabrica una oblea.
- Definir los diferentes tipos y técnicas de crecimiento epitaxial.
- Presentar los principales defectos que se pueden producir en un cristal y comentar sus efectos en el comportamiento eléctrico del semiconductor.

Palabras clave

Método de Czochralski.
Método de Zona Flotante.
Policristalin
Monocristalino.
Coeficiente de segregación.
Técnica LEC.
Método de Bridgman.
Homoepitaxia.
Heteroepitaxia.
Epitaxia en fase liquida o LPE
Epitaxia en fase gaseosa o VPE
Epitaxia por haces moleculares (M.B.E.).
Posición intersticial
Posición substitucional
Defecto Frenkel.
Gettering.

9.1 Introducción

Los semiconductores más importantes para la fabricación tanto de dispositivos discretos como de circuitos integrados son, con diferencia, el silicio (Si) y el arseniuro de galio (GaAs). El diagrama de la **Figura 9.1.1** muestra el proceso de crecimiento de una oblea del semiconductor correspondiente a partir de los materiales iniciales:

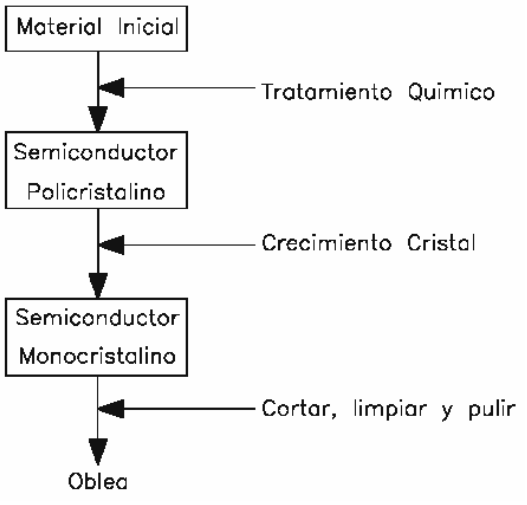


Figura 9.1.1: Esquema del proceso de fabricación de una oblea.

Los materiales de partida (SiO_2 para el silicio y galio y arsénico para el GaAs) son tratados químicamente para obtener un semiconductor policristalino de gran pureza. Partiendo de este semiconductor de alta pureza, se crecen, por diferentes técnicas lingotes de semiconductor monocristalino de determinado diámetro. Finalmente estos lingotes se cortan en obleas que son limpiadas y pulidas. Sobre estas superficies especulares se fabricarán por último los diferentes dispositivos electrónicos. Una tecnología relacionada muy de cerca con el crecimiento de cristales, involucra el crecimiento de capas delgadas de semiconductor monocristalino. Esta técnica se denomina Epitaxia. El crecimiento epitaxial permite controlar los perfiles de dopado, optimizando así la fabricación de dispositivos y circuitos.

9.2 Crecimiento en volumen. Obtención de Si cristalino.

9.2.1 Obtención de Si puro.

El material inicial para el crecimiento de lingotes de silicio es una forma relativamente pura de arena (SiO_2) denominada cuarcita. La cuarcita se coloca en un horno junto con varias formas de carbón (hulla, coke, astillas de madera), dando lugar a la reacción siguiente:

$$\mathbf{SiC}(\text{solido}) + \mathbf{SiO_2}(\text{solido}) \rightarrow \mathbf{Si}(\text{solido}) + \mathbf{SiO}(\text{gas}) + \mathbf{CO(}\text{gas})$$

Esta reacción produce silicio metalúrgico (MGS) con una pureza del 98%. Este silicio no es todavía lo suficientemente puro para poder utilizarlo en la fabricación de circuitos electrónicos. Por lo tanto es necesario un proceso de purificación. Para llevar a cabo tal proceso, el silicio es pulverizado y tratado con cloruro de hidrógeno para obtener triclorosilano ($SiHCl_3$), de acuerdo con la reacción:

$$\mathbf{SiO_2(solido) + 3HCl(gas) \rightarrow SiHCl_3(gas) + H_2(gas)}$$

A temperatura ambiente el triclorosilano es un líquido. La destilación fraccionada de este líquido permite eliminar las impurezas indeseadas. A continuación, el triclorosilano se reduce con hidrógeno para obtener silicio electrónico (EGS : Electronic Grade Silicon):

$$\mathbf{SiHCl_3}(\text{gas}) + \mathbf{H_2}(\text{gas}) \rightarrow \mathbf{Si}(\text{solido}) + \mathbf{3HCl}(\text{gas})$$

Esta reacción tiene lugar en un reactor que contiene una barra de silicio caliente que sirve para que el silicio electrónico se deposite sobre ella. El EGS es un silicio policristalino de alta pureza (concentración de impurezas en una parte por mil millones) y es el elemento de partida para crear silicio monocristalino.

9.2.2 Crecimiento en volumen.

Una vez que se ha conseguido silicio de alta pureza o EGS (Electronic Grade Silicon). Para la fabricación de un CI se requiere Silicio con estructura cristalina. Para conseguir un cristal de Si se pueden utilizar varias técnicas. Las más importantes son: el método de Czochralski y el método de Zona Flotante

Metodo de Czochralski

El método de Czochralski es el método empleado en el 90% de los casos para obtener silicio monocristalino a partir de silicio policristalino (EGS). Este método utiliza para el crecimiento de cristales un aparato denominado "puller", que consta de tres componentes principales como muestra la Figura 9.2.1

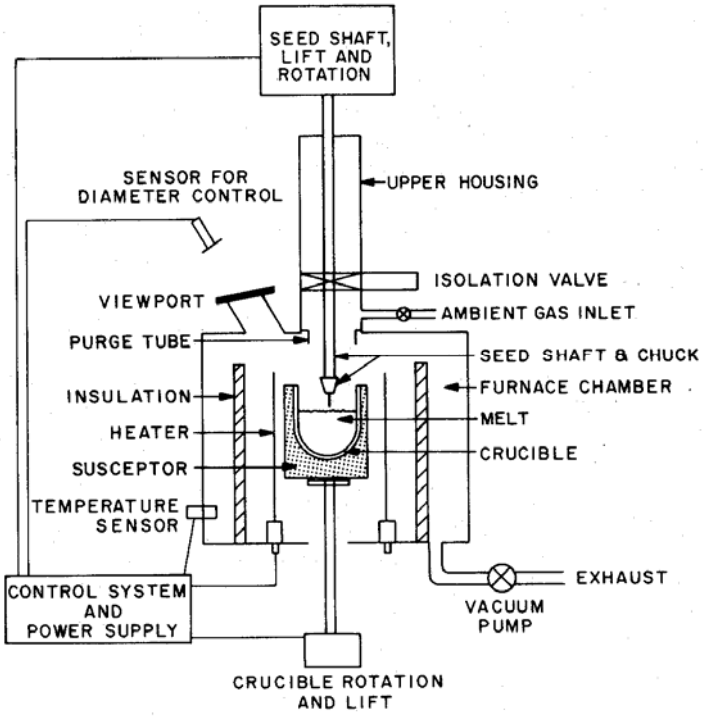


Figura 9.2.1 Puller de Czochralski. Fuente:Libro VLSI Technology de Sze

(a) Un horno, que incluye un crisol de sílice fundida (SiO_2), un soporte de grafito, un mecanismo de rotación (en el sentido de las agujas del reloj) un calentador y una fuente de alimentación.

(b) Mecanismo de crecimiento del cristal, que incluye un soporte para la semilla (muestra patrón del cristal que se pretende crecer) y un mecanismo de rotación (en el sentido contrario al de las agujas del reloj).

(c) Mecanismo del control de ambiente. Incluye una fuente gaseosa (argón por ejemplo), un mecanismo para controlar el flujo gaseoso y un sistema de vaciado.

El proceso de crecimiento se detalla a continuación. El silicio policristalino (EGS) se coloca en el crisol y el horno se calienta a una temperatura superior a la de fusión del silicio obteniéndose el material fundido (MELT). A continuación, se suspende sobre el crisol una muestra pequeña del tipo de cristal que se quiere crecer (<111> por ejemplo). Al introducir la semilla en el fundido, parte de la misma se funde, pero la punta de la misma aún toca a la superficie del líquido. Entonces lentamente empieza a levantarse. El progresivo enfriamiento en la interfase sólido-líquido proporciona un silicio monocristalino con la misma orientación cristalina que la semilla pero de mayor diámetro.

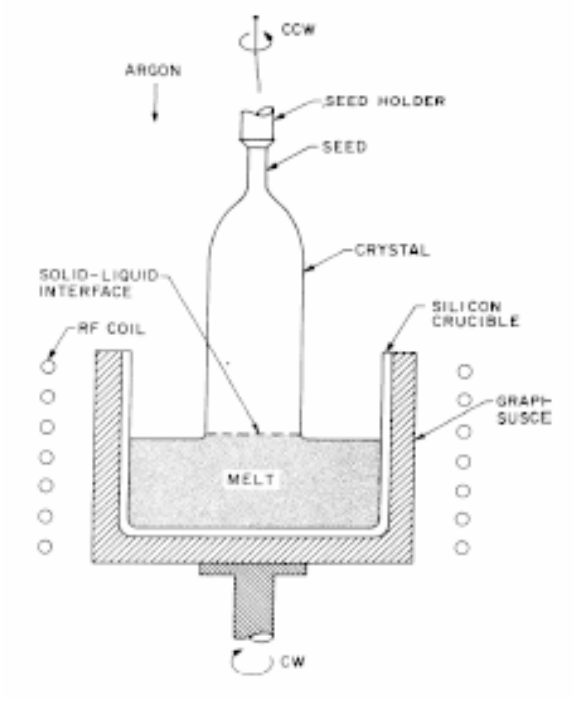


Figura 9.2.2: Formación de una barra de Si cristalino método de Czochralski. Fuente:Libro VLSI Technology de Sze

Controlando cuidadosamente la temperatura, la velocidad de elevación y rotación de la semilla y la velocidad de rotación del crisol, se mantiene un diámetro preciso del cristal. Mientras los lingotes son estirados, se refrescan para que adquieran un estado sólido. La longitud del lingote vendrá determinada por la cantidad de silicio fundido que hay en el crisol. La velocidad típica de crecimiento es de pocos milímetros por minuto.

En este proceso se añaden la cantidad de impurezas necesarias para formar un semiconductor tipo N o P con el dopado deseado. Normalmente la concentración de impurezas es de 10^{15} cm^{-3}. Para conseguir esta concentración se incorpora cuidadosamente una pequeña cantidad de dopante por ejemplo Fósforo (para conseguir semiconductor tipo N) o Boro (para tipo P) al Silicio fundido. La cantidad típica es de un microgramo por cada kilogramos de silicio. Para tener un mejor control se añade silicio altamente dopado al Silicio intrínseco fundido.

La concentración de dopante del silicio una vez que se solidifica es siempre inferior a la del silicio fundido. Esta segregación causa que la concentración del dopante aumente a medida que la barra de cristal crece. La concentración de impurezas es menor en lado de la semilla que en el otro extremo. También se tiene un pequeño gradiente de concentración a lo largo del radio de la barra de cristal. La relación entre estas dos concentraciones se define como *coeficiente de segregación en equilibrio, k_0:*

$$k_0 \equiv \frac{C_s}{C_l}$$

donde C_s y C_l son las concentraciones en equilibrio en el sólido y en el líquido respectivamente.

La **Figura 9.2.3** muestra algunos lingotes crecidos por el método de Czochralski.

Figura 9.2.3: Lingotes de Si crecidos por el método de Czochralski
http://www.sumcosi.com/products/products2.html

Como se observa en la tabla siguiente, el valor de k_0 para la mayoría de los dopantes habitualmente utilizados es menor que la unidad lo que significa que durante el crecimiento los dopantes son rechazados hacia el fundido. Consecuentemente, éste estará progresivamente más enriquecido en dopante a medida que se crece el cristal.

Si			GaAs		
Dopante	k_0	Tipo	Dopante	k_0	Tipo
As	0.3	n	S	0.5	n
Bi	$7x10^{-4}$	n	Se	0.1	n
C	0.07	n	Sn	0.08	n
Li	10^{-2}	n	Te	0.064	n
O	0.5	n	C	1	n/p
P	0.35	n	Ge	0.018	n/p
Sb	0.023	n	Si	2	n/p
Te	$2x10^{-4}$	n	Be	3	p
Al	$2.8x10^{-3}$	p	Mg	0.1	p
Ga	$8x10^{-3}$	p	Zn	0.42	p
B	0.8	p	Cr	$5.7x10^{-4}$	Semi-aislante
Au	$2.5x10^{-5}$	Deep lying	Fe	$3x10^{-3}$	Semi-aislante

Tabla 9.2.1: Coeficientes de segregación para distintos dopantes

Consideremos el caso de un cristal que está siendo crecido a partir de un fundido de masa inicial M_0 y concentración de dopante inicial M_0. Sea S la cantidad de dopante que queda en el fundido cuando se ha crecido un cristal de masa M. Supongamos ahora que el cristal aumenta su masa en dM, lo que corresponde a una disminución -dS en el peso del dopante presente en el fundido.

Si C_s es la concentración de impurezas en el cristal en ese momento, $C_s dM$ será el peso de las impurezas añadidas al cristal en este crecimiento infinitesimal, y en consecuencia

$$-dS = C_s\, dM \tag{9.1}$$

El peso del fundido restante tras crecer un cristal de masa M, será M_0-M, por lo que la concentración de dopante en el líquido es:

$$C_l = \frac{S}{M_0 - M} \tag{9.2}$$

de donde,

$$-dS = C_l\, k_0\, dM = k_0 \frac{S\, dM}{M_0 - M} \tag{9.3}$$

obteniéndose finalmente la siguiente ecuación diferencial:

$$\frac{dS}{S} = -k_0 \frac{dm}{M_0 - M} \tag{9.4}$$

Conocida la cantidad inicial de dopante presente en el fundido, $S_0=C_0M_0$, podemos integrar la ecuación anterior:

$$\int_{C_0M_0}^{S} \frac{dS}{S} = k_0 \int_0^M \frac{-dM}{M_0 - M} \tag{9.5}$$

Integrando la ecuación anterior obtenemos:

$$\frac{S}{C_0 M_0} = \left(1 - \frac{M}{M_0}\right)^{k_0} \tag{9.6}$$

Como por otro lado tenemos que

$$C_l = \frac{S}{M_0 - M} \tag{9.7}$$

$$\frac{C_l (M_0 - M)}{C_0 M_0} = \left(1 - \frac{M}{M_0}\right)^{k_0} \tag{9.8}$$

De donde finalmente se obtiene la concentración de dopante en el cristal:

$$C_s = C_0 k_0 \left(1 - \frac{M}{M_0}\right)^{k_0 - 1} \tag{9.9}$$

A medida que el cristal crece, la concentración inicial de dopado C_0k_0 va disminuyendo si $k_0<1$, e irá aumentando para $k_0>1$. Cuando $k_0 \bullet 1$, se obtiene una distribución uniforme de impurezas.

Coeficiente de segregación efectivo

Mientras se crece el cristal, los dopantes son rechazados hacia el fundido continuamente ($k_0<1$). Si la velocidad de rechazo es mayor que la velocidad a la que el dopante es transportado por difusión hacia el interior del fundido, se produce en el líquido un gradiente de concentración de dopante, como muestra Figura 9.2.4. Podemos definir un coeficiente de segregación efectivo, k_e, cómo la relación entre la concentración de dopante en el sólido C_s, y la concentración de dopante en el fluido lejos de la interface, C_l.

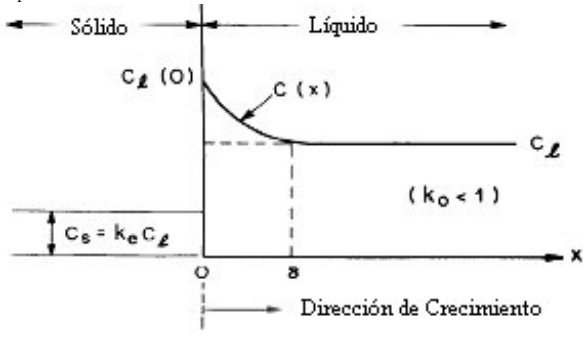


Figura 9.2.4: Gradiente de concentración de dopante

Consideremos una pequeña lámina de fundido de espesor δ, tal que fuera de esta lámina la concentración de dopante tiene una concentración constante C_l. Dentro de la lámina, la concentración de dopante puede describirse, en condiciones estacionarias, por la ecuación de continuidad (analogía con la ecuación de continuidad de portadores en semiconductores):

$$0 = v\frac{dC}{dx} + D\frac{d^2C}{dx^2} \qquad (9.10)$$

donde v es la velocidad de crecimiento del cristal, D el coeficiente de difusión del dopante en el fundido y C la concentración de dopante en el fundido. La solución a esta ecuación diferencial de 2 orden con coeficientes constantes es de la forma:

$$C = A_1 e^{\frac{v}{D}x} + A_2 \qquad (9.11)$$

donde A_1 y A_2 son las constantes a determinar por las condiciones de contorno. La primera condición de contorno es que $C=C_l(0)$ en x=0. La segunda condición es la de conservación del número total de dopantes, esto es, la suma de flujos de dopante en la interface debe ser cero. Considerando la difusión de los átomos de dopante en el fundido (despreciando la difusión en el sólido) tendremos que:

$$D\left(\frac{dC}{dx}\right)_{x=0} + \left(C_l(0) - C_s\right)v = 0 \qquad (9.12)$$

Teniendo en cuenta las condiciones de contorno anteriores y que $C=C_l$ para $x=\delta$, tendremos que

$$e^{\frac{v\delta}{D}} = \frac{C_l - C_s}{C_l(0) - C_s} \tag{9.13}$$

por lo que el coeficiente de segregación efectiva queda en la forma:

$$k_e \equiv \frac{C_s}{C_l} = \frac{k_0}{k_0 + (1 - k_0)e^{\frac{v\delta}{D}}} \tag{9.14}$$

mientras que la distribución de dopante en el cristal sigue la misma ley anterior, pero sustituyendo k_0 por k_e:

$$C_s = C_0 k_e \left(1 - \frac{M}{M_0}\right)^{k_e - 1} \tag{9.15}$$

Los valores de k_e son mayores que los de k_0 y puede aproximarse a 1 para valores grandes del parámetro de crecimiento $v\delta/D$. La distribución uniforme (k_e61) puede obtenerse empleando una velocidad elevada de crecimiento y una velocidad de rotación pequeña (δ es inversamente proporcional a la velocidad de rotación).

El Silicio fabricado por el método de Czochralski contiene una considerable cantidad de oxigeno, debido a la disolución del crisol de Sílice (SiO_2). Este oxigeno no es perjudicial para el silicio de baja resistividad usado en un circuito integrado, además puede controlar el movimiento accidental de impurezas metálicas. Sin embargo para aplicaciones de alta potencia donde se necesita Si con alta resistividad este oxigeno es un problema. En estos casos se usa el método de Zona Flotante.

Metodo de Zona flotante

El método de "float-zone" se utiliza para crecer silicio monocristalino con concentración de impurezas más bajas que las normalmente obtenidas por el método de Czochralski.

El cilindro de silicio policristalino se sostiene verticalmente y se conecta a uno de sus extremos a la semilla, girándose todo como muestra la figura. El cilindro de silicio se encierra en un recipiente de cuarzo y se mantiene en una atmósfera inerte (argón por ejemplo).

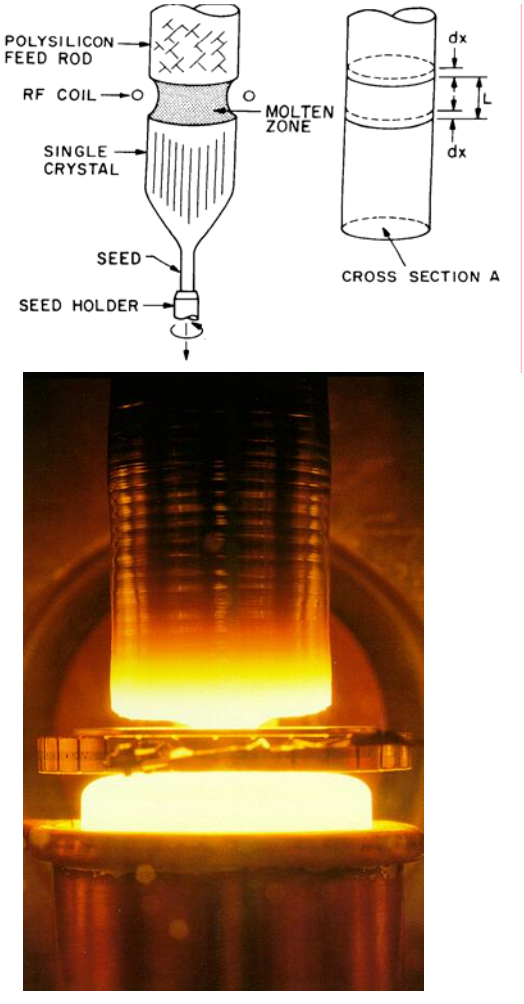


Figura 9.2.5: Método de zona flotante
http://www.oja-services.nl/iea-pvps/ ar02/dnk.htm

Durante la operación, una pequeña zona (pocos centímetros) del cristal se hunde mediante un calentador que se desplaza a lo largo de todo el cristal desde la semilla. El silicio fundido es retenido por la tensión superficial entre ambas caras del silicio sólido. Cuando la zona flotante se desplaza hacia arriba, el silicio monocristalino se solidifica en el extremo inferior de la zona flotante y crece como una extensión de la semilla. Mediante este proceso de "float zone" pueden obtenerse materiales con resistividades más altas que mediante el método de Czochralski. Además, como no se necesita crisol, no existe, como en el caso anterior, posible contaminación desde el crisol.

Figura 9.2.6: Puller para zona flotante de la empresa GTI
http://www.gt-equipment.com/

Para evaluar la distribución de dopado de un proceso "float-zone", consideremos el modelo simplificado de la figura anterior. Supongamos que la concentración inicial uniforme de dopante en el cilindro policristalino es C_0. Sea L la longitud de la zona fundida a una distancia x, A la sección transversal del cilindro y ρ_d la densidad del silicio. Sea S la cantidad de dopante presente en la zona fundida. Cuando la zona fundida avanza dx, la cantidad de dopante añadida al fundido es:

$$C_0 \rho_d A\, dx \tag{9.16}$$

mientras que la cantidad de dopante que desaparece debido a que es incorporada al silicio monocristalino es:

$$C_s \rho_d A\, dx = k_e \left(S \frac{dx}{L} \right) \tag{9.17}$$

donde k_e es el coeficiente de segregación efectiva. En consecuencia la variación en la cantidad de dopante en la zona fundida es:

$$dS = \left(C_0 \rho_d A - \frac{k_e S}{L} \right) dx \tag{9.18}$$

Integrando

$$\int_0^x dx = \int_{S_0}^{S} \frac{dS}{C_0 \rho_d A - \frac{k_e S}{L}} \tag{9.19}$$

donde $S_0=C_0\rho_d AL$ es la cantidad inicial de dopante en la zona fundida. Como resultado de la integral anterior se obtiene:

$$S = \frac{C_0 A \rho_d L}{k_e} \left[1 - (1 - k_e)\, e^{-\frac{k_e x}{L}} \right] \tag{9.20}$$

Como $C_s=k_eC_l$, y C_l que es la concentración de dopante en el líquido viene dada por:

$$C_l = \frac{S}{A \rho_d L} \tag{9.21}$$

finalmente la concentración de dopante en el cristal se expresa en la forma:

$$C_s = C_0 \left[1 - (1 - k_e)\, e^{\frac{k_e x}{L}}\right] \qquad (9.22)$$

Puesto que C_s es menor que C_0, este método, al igual que el método de Czochralski puede emplearse para purificar cristales.

Para ciertos dispositivos conmutadores (tiristores) se utilizan grandes áreas del chip (una oblea entera para un sólo dispositivo). Esto implica la necesidad de un dopado homogéneo en toda la oblea. Para obtener esto, se utiliza una lámina de silicio monocristalino crecida mediante la técnica de "float-zone" con un dopado mucho más pequeño que el finalmente requerido. A continuación, la lámina se irradia con neutrones térmicos. Este proceso denominado irradiación de neutrones da lugar a una transmutación del silicio en fósforo obteniéndose silicio monocristalino tipo-n:

$$Si_{14}^{30} + \text{neutrones} \rightarrow Si_{14}^{31} + \text{radiaci n } \gamma \rightarrow P_{15}^{31} + \text{radiaci n } \beta$$

La profundidad de penetración de neutrones en silicio es de 100cm, por lo que el dopado es muy uniforme en toda la lámina.

9.3 Crecimiento de materiales compuestos. Técnica LEC. Método de Bridgman.

Los materiales iniciales para la obtención de arseniuro de Galio (GaAs) son los elementos puros arsénico y galio. Estos elementos se utilizan para sintetizar GaAs policristalino. Al ser el arseniuro de galio una combinación de dos materiales, su comportamiento es muy diferente al del silicio. La figura siguiente muestra el diagrama de fases del sistema galio-arsénico:

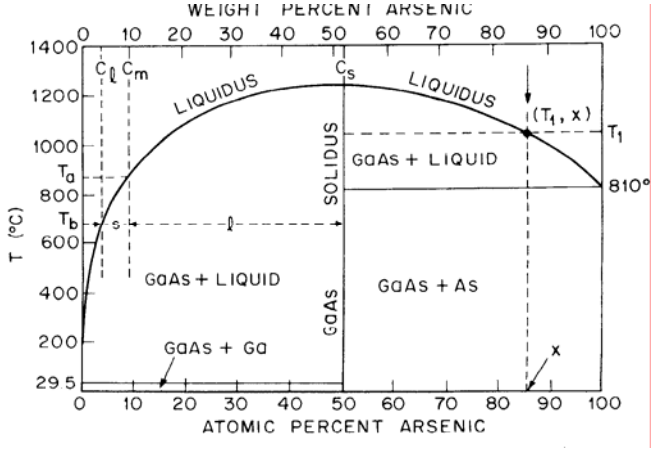


Figura 9.3.1: Diagrama de fases galio-arsénico.

La abcisa representa la composición de los dos componentes en términos del porcentaje de peso atómicos (escala inferior) y porcentaje de peso (escala superior). Consideremos por ejemplo, un fundido con una composición en arsénico inicial x. Cuando se baja la temperatura, la composición permanecerá fija hasta que se alcance la línea "líquidus". Cuando se alcance dicha temperatura, (T_1,x), empezará a solidificar material con un 50% de átomos de arsénico, es decir, arseniuro de galio.

Al contrario que el silicio, el arsénico y el galio tienen presiones de vapor muy elevadas a la temperatura de fusión del arseniuro de galio. Esto significa que el arsénico es volátil y se evapora fácilmente a la temperatura de fusión, tendiendo a abandonar el fluido y a condensarse sobre las paredes del recipiente. En su fase de vapor, el arsénico existe cómo As_2 y As_4. Para evitar la salida del arsénico del fundido, todas las manipulaciones de éste deben realizarse en condiciones de sobre-presión de arsénico.

Para sintetizar arseniuro de galio, se utiliza un sistema que consiste en un tubo de cuarzo sellado en el que se ha hecho el vacío. Este tubo se coloca en un horno tal que tiene dos zonas, cada una de ellas a una temperatura diferente. El arsénico de alta pureza se coloca en un recipiente de grafito y se calienta a 610 C, mientras que el galio de alta pureza se coloca en otro recipiente de grafito y se calienta hasta por encima de la temperatura de fusión del arseniuro de galio (1238 C) En estas condiciones, se establece una sobrepresión de arsénico que:

(1).-permite el transporte de vapor de arsénico hacia el fundido de galio, obteniéndose arseniuro de galio, y

(2).-evita la descomposición del arseniuro de galio cuando se forma (debido a la alta presión de vapor de arsénico).

Cuando el fundido se enfría, se obtiene arseniuro de galio policristalino de alta pureza.

9.3.2 Técnica LEC (Liquid-Encapsulated-Czochralski).

Para obtenerse GaAs monocristalino a partir del semiconductor policristalino, pueden emplearse diferentes técnicas. Puede utilizarse el método de Czochralski anteriormente detallado para el silicio. Sin embargo, debido a la alta presión de vapor del arsénico a la temperatura de fusión del arseniuro de galio, hay que tomar ciertas precauciones puesto que de lo contrario, el arsénico abandonaría rápidamente el fundido con lo que únicamente quedaría galio en el fundido. Por ejemplo, el proceso puede hacerse en un ambiente con alta presión de arsénico, para evitar la evaporación del arsénico presente en el fundido. Sin embargo, lo que realmente se hace es aislar el fundido mediante una capa de un segundo material (óxido bórico) que impide la salida del arsénico del fundido como muestra la **Figura 9.3.2**

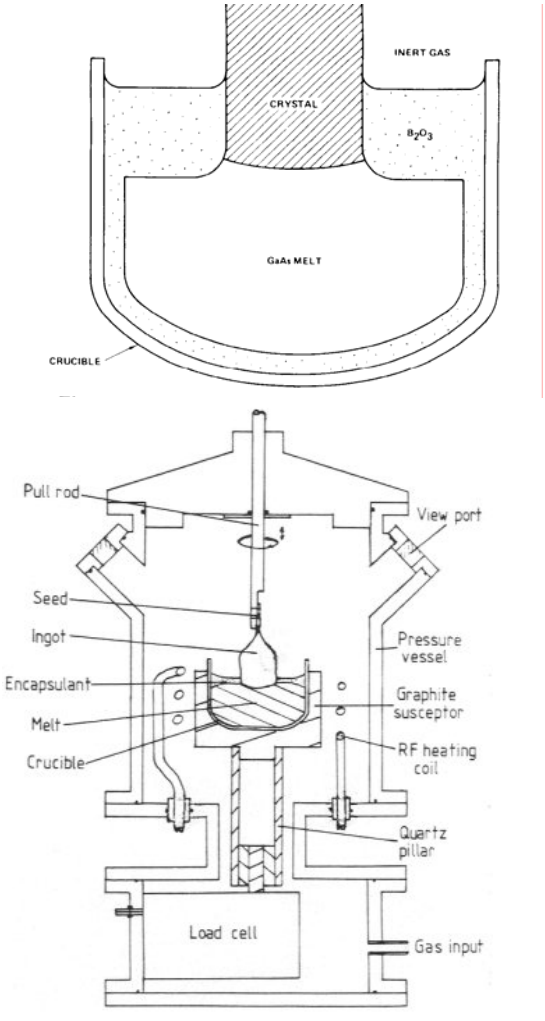


Figura 9.3.2: (a) Técnica LEC, (b) Puller Czochralski
http://www.taloma.com/cwarwick/ thesis/figure_3_1.gif

A esta técnica se le denomina técnica LEC (Liquid-Encapsulated-Czochralski), y permite, de la misma forma que el método de Czochralski para el silicio, obtener GaAs monocristalino.

9.3.3 Técnica de Bridgman

Existen otras técnicas que permiten obtener GaAs monocristalino a partir de un fundido de GaAs policristalino. La figura siguiente muestra un sistema que emplea la técnica de Bridgman para obtener GaAs monocristalino:

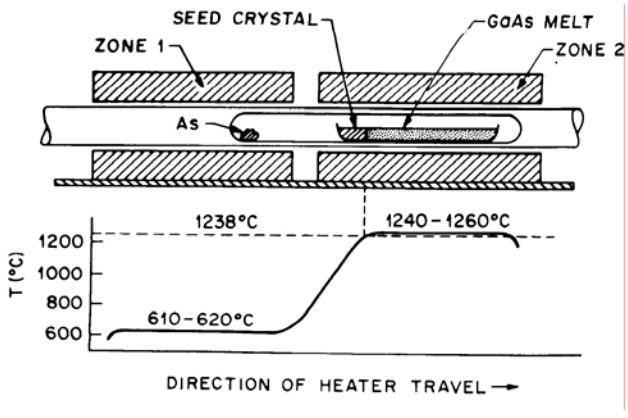


Figura 9.3.3: Técnica de Bridgman

Se utiliza un horno de dos zonas. La zona de la izquierda se mantiene a una temperatura de 610 C para producir en el sistema una sobre presión de arsénico que impida la vaporización excesiva del arsénico en el GaAs fundido. La zona derecha, al contrario, se mantiene a una temperatura justo por encima del punto de fusión del arseniuro de galio. El tubo sellado está fabricado de cuarzo, y el recipiente donde se coloca el GaAs policristalino de grafito. Cuando el horno se desplaza hacia la derecha, el GaAs fundido se enfría por el extremo izquierdo, donde se coloca una semilla (patrón con la orientación cristalográfica deseada para el cristal resultante). El enfriamiento gradual del fundido de esta forma permite el crecimiento de un cristal de GaAs. La distribución de impurezas puede describirse por las mismas ecuaciones anteriores:

$$C_s = C_0\, k_e \left(1 - \frac{M}{M_0}\right)^{k_e - 1} \tag{9.23}$$

$$k_e \equiv \frac{C_s}{C_l} = \frac{k_0}{k_0 + (1 - k_0)\, e^{-\frac{v\delta}{D}}} \tag{9.24}$$

donde la velocidad de crecimiento viene dada por la velocidad de desplazamiento del horno en la dirección perpendicular.

Figura 9.3.4: Horno vertical de Bridgman de tres zonas (izquierda) : Horno vertical de Bridgman de cuatro zonas industrial (derecha) http://www.cyberstar.fr/crystal/bridgman.htm

9.4 Producción de obleas.

Figura 9.4.1: Lingote monocristalino de Czochralski

Después de crecido el cristal, la primera operación a realizar es quitar los extremos del lingote, tanto el de la semilla, como el último extremo crecido:

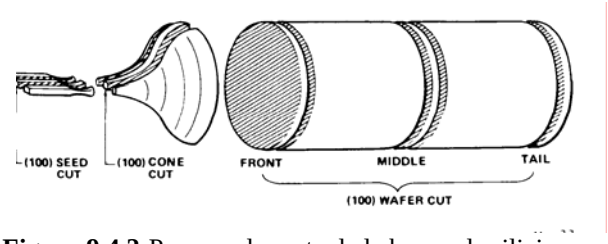


Figura 9.4.2: Proceso de corte de la barrra de silicio.

La operación siguiente es desgastar la superficie hasta que quede definido el diámetro del lingote. A continuación, y paralela a la generatriz del cilindro se hacen unas marcas planas para especificar la orientación del cristal y el tipo de conductividad del material. La figura muestra las marcas realizadas y el significado de éstas:

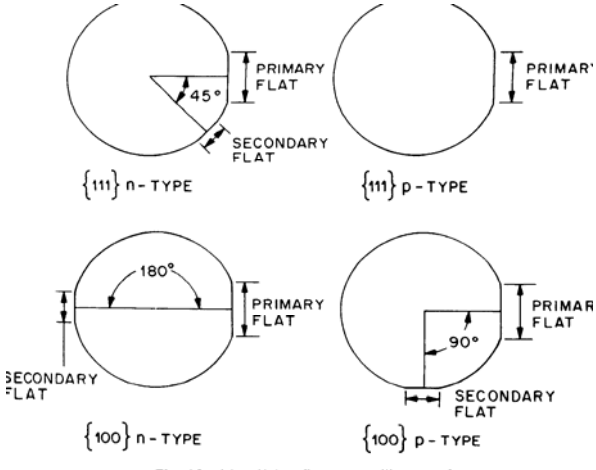


Figura 9.4.3: Tipos de marcas y su significado.

Una vez realizadas estas operaciones, el lingote está preparado para ser cortado en obleas. Cortadas las obleas, las dos caras de estas son tratadas con una mezcla de Al_2O_3 y glicerina para producir una superficie plana homogénea con un error de "2 μm. Esta operación daña y contamina la superficie y bordes de la oblea. Para reparar estos daños, las obleas son tratadas mediante ataques químicos que posteriormente veremos. El paso final en la obtención de las obleas es el pulido, cuyo propósito es obtener una superficie especular dónde puedan definirse los detalles de los dispositivos electrónicos.

9.5 Crecimiento epitaxial.

El término epitaxia procede del griego "epi" (sobre) y taxis (arreglo), y se aplica al proceso usado para crecer una capa delgada cristalina sobre un substrato cristalino. En un proceso epitaxial, el substrato de la oblea actúa como la semilla en el crecimiento de un cristal. Los procesos epitaxiales se diferencian de los procesos de crecimiento de volumen antes mencionados en que la capa epitaxial puede crecerse a temperaturas substancialmente más pequeñas que las del punto de fusión del material (sobre un 30-50 % más bajas). Cuando un material se crece epitaxialmente sobre un substrato del mismo material, el proceso se denomina homoepitaxia. Si la capa y el substrato son de materiales diferentes, tal como el caso de AlxGa1-xAs sobre GaAs, el proceso se denomina heteroepitaxia. En cualquier caso, en heteroepitaxia, las estructuras cristalinas del substrato y de la capa crecida deben ser parecidas si se pretende obtener un crecimiento cristalino.

Existen diferentes tipos de procesos epitaxiales:

-Epitaxia en fase gaseosa o VPE, en la que se crece la capa cristalina a partir de los reactivos en fase gaseosa.

-Epitaxia en fase líquida ó LPE, en la que los reactivos utilizados para crecer la capa cristalina están en fase líquida.

-Epitaxia por haces moleculares ó MBE. Los reactivos involucrados en este proceso son haces de átomos o moléculas, en un entorno de muy alto vacío.

En los dispositivos de silicio, el proceso epitaxial más importante es la epitaxia en fase gaseosa (VPE):

9.5.2 Vapour Phase Epitaxy (VPE)

Las obleas de silicio se introducen en un recipiente sobre un soporte de grafito, como muestra la figura:

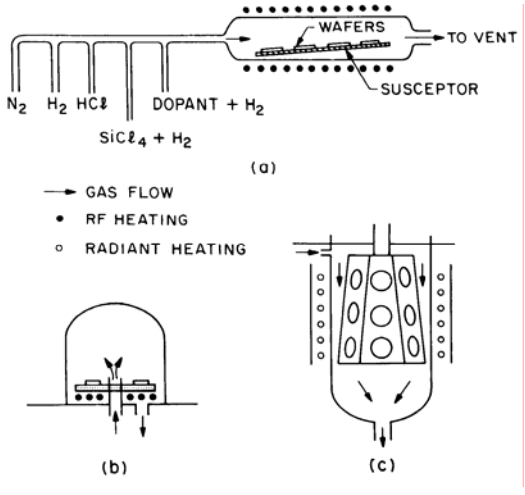


Figura 9.5.1: Vapour-Phase Epitxy (VPE)

En el recipiente se introduce la fuente gaseosa, típicamente tetracloruro de silicio ($SiCl_4$) y se calienta todo a una temperatura de 1200 C, dándose la reacción:

$$SiCl_4(gas) + 2H_2(gas) \Rightarrow Si(solido) + 4HCl(gas)$$

Pero además se produce también la reacción siguiente:

$$SiCl_4(gas) + Si(solido) \Rightarrow 2SiCl_2(gas)$$

Por lo tanto, si la concentración de tetracloruro de silicio ($SiCl_4$) es demasiado elevada, predominará la segunda reacción, por lo que se producirá una eliminación de silicio del substrato en vez del crecimiento de la capa epitaxial.

La **Figura 9.5.2**muestra el efecto de la concentración de $SiCl_4$ en el gas reactivo sobre la velocidad de crecimiento de la capa epitaxial. Inicialmente, la velocidad crecimiento aumenta linealmente con la concentración de $SiCl_4$, alcanzándose un valor máximo. Después de este máximo, la velocidad de crecimiento disminuye a medida que aumenta la concentración de $SiCl_4$, llegando un momento en que la velocidad es negativa, es decir, ocurre la reacción de eliminación de silicio. Generalmente la capa epitaxial de silicio se crece en la región de bajas concentraciones en la que se verifica un comportamiento lineal:

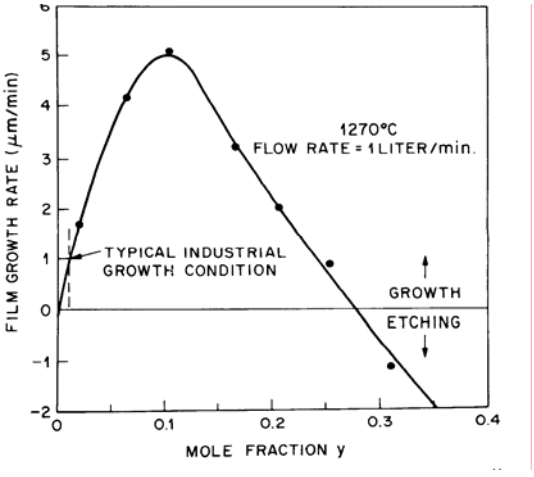


Figura 9.5.2: Velocidad de crecimiento (VPE)

La capa epitaxial puede crecerse con un cierto dopado. El dopante se introduce a la vez que el $SiCl_4$ en la mezcla gaseosa. Como dopante tipo p se utiliza el diborano (B_2Cl_4), mientras que la arsina (AsH_3) y la fosfina (PH_3) se utilizan como dopantes tipo n.

Es posible crecer epitaxialmente también arseniuro de galio mediante un proceso de epitaxia en fase gaseosa. Como el arseniuro de galio se descompone en arsénico y galio por evaporación, no es posible su transporte directo en fase de vapor. Un procedimiento alternativo es usar As_4 para el componente de arsénico y cloruro de galio $GaCl_3$ para la componente de galio. La reacción que conduce al crecimiento epitaxial del GaAs es:

$$As_4 + 4GaCl_3 + 6H_2 \Rightarrow 4GaAs + 12HCl$$

El As_4 se genera térmicamente por descomposición de la arsina:

$$As_4 + 4GaCl_3 + 6H_2 \Rightarrow 4GaAs + 12HCl$$

El As_4 se genera térmicamente por descomposición de la arsina:

$$4\,AsH_3 \rightarrow As_4 + 6\,H_2$$

mientras que el cloruro de galio se obtiene a través de la reacción:

$$6\,HCl + 2\,Ga \rightarrow 2\,GaCl_3 + 3\,H_2$$

Los reactivos se introducen en el reactor con un gas portador (H_2), mientras que las obleas de GaAs se mantienen a una temperatura entre 650 C y 850 C. Para evitar la descomposición térmica del substrato y de la lámina crecida cómo consecuencia de la elevada presión de vapor del arsénico, debe existir una sobrepresión suficiente de arsénico.

Otro procedimiento alternativo para obtener una capa epitaxial de GaAs es el proceso de deposición química de gases metal-orgánicos: **MOCVD** (Metal-Organic-Chemical-Vapor-Deposition). Se utilizan compuestos metalorgánicos como el trimetilgalio, $Ga(CH_3)_3$, que junto con la arsina (AsH_3) proporcionan, de acuerdo con la reacción siguiente la capa epitaxial de GaAs:

$$AsH_3 + Ga(CH_3)_3 \rightarrow GaAs + 3CH_4$$

Durante el proceso de epitaxia, puede introducirse la cantidad adecuada de dopante en el arseniuro de galio para obtener el dopado

deseado. El proceso se realiza añadiendo los dopantes en fase gaseosa a la mezcla gaseosa utilizada. Como dopantes tipo n, se utilizan hidruros de azufre o selenio, o tetrametilestaño. Para conseguir un dopado tipo p, se añade dietilzinc o dietilcadmio; también se utiliza cloruro de cromilo (CrO_2Cl_2) para introducir cromo en el arseniuro de galio para obtener capas semiaislantes.

-Cinética de crecimiento.

Consideremos la figura siguiente:

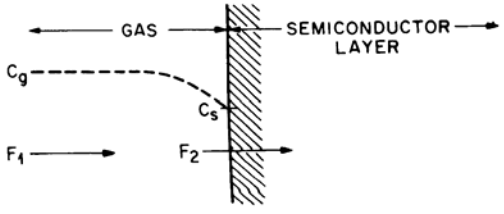


Figura 9.5.3: Esquema de una reacción VPE

Sea C_g la concentración de reactivos en la mezcla gaseosa, lejos de la interface gas-substrato. Sea C_s la concentración de reactivos justo en la interface. Sea F_1 el flujo de reactivos desde el volumen del gas hacia la interface (flujo es el número de moléculas que cruzan un área unidad en la unidad de tiempo). Sea F_2 el flujo correspondiente de especies reactivas consumidas en la reacción epitaxial. Supongamos que F_1 viene dado por:

$$F_1 = h_g (C_g - C_s) \tag{9.25}$$

donde h_g es el coeficiente de transferencia de masa en la fase de vapor, y viene dada por:

$$h_g = \frac{3}{2} D_g \sqrt{\frac{\rho_d v}{\mu L}} \tag{9.26}$$

siendo D_g la difusividad del gas, μ su viscosidad, ρ_d su densidad, v la velocidad a la que es introducido el gas y L la longitud del tubo donde se realiza el crecimiento epitaxial.

Por otro lado, el flujo consumido por la reacción química que tiene lugar en la superficie de la capa crecida puede expresarse cómo:

$$F_2 = k_s C_s \tag{9.27}$$

donde k_s es la constante de velocidad de la reacción. En el estado estacionario ambos flujos serán idénticos, por lo que tendremos que:

$$F \equiv F_1 = F_2 \tag{9.28}$$

de donde se obtiene que la concentración de reactivos en la interface es:

$$C_s = \frac{C_g}{1 + \left(\frac{k_s}{h_g}\right)} \qquad (9.29)$$

La velocidad de crecimiento de la capa semiconductora viene dada por el flujo de estados estacionarios, F, dividido por el número de átomos de semiconductor incorporados a una unidad de volumen de la lámina, C_a:

$$v = \frac{F}{C_a} = \frac{k_s\, h_g}{k_s + h_g}\left(\frac{C_g}{C_a}\right) \qquad (9.30)$$

C_a tiene un valor de $5x10^{22}$ átomos/cm^3 para el silicio y de $4.4x10^{22}$ átomos/cm^3 para el arseniuro de galio.

Cómo C_g= y C_t, donde "y" es la fracción molar de las especias reactivas y C_t el número total moléculas por cm^3 de gas (reactivos más gas portador) se obtiene que la velocidad de crecimiento de la capa epitaxial:

$$v = \frac{F}{C_a} = \frac{k_s\, h_g}{k_s + h_g}\left(\frac{C_g}{C_a}\right) \qquad (9.31)$$

La expresión anterior indica que la velocidad de crecimiento, para una fracción molar dada, está determinada por el menor de k_s o h_g:

- Si k_s n h_g entonces la velocidad de crecimiento viene dada por:

$$k_s \gg h_g \;\_\; v = k_s\left(\frac{C_t}{C_a}\right) y \qquad (9.32)$$

es decir, la velocidad de crecimiento está determinada por cómo de rápido se produzca la reacción en la superficie del semiconductor. Se dice que el proceso está controlado por la velocidad de reacción superficial.

- Por otro lado, si k_s o h_g, entonces, la velocidad de crecimiento:

$$k_s \ll h_g \;\_\; v = h_g\left(\frac{C_t}{C_a}\right) y \qquad (9.33)$$

es decir, la velocidad de crecimiento está determinada por cómo de rápido llegan los reactivos a la superficie de la oblea. En este caso se dice que el proceso está controlado por la transferencia de masa.

La **Figura 9.5.4**muestra la dependencia con la temperatura de la velocidad de crecimiento para varias fuentes de silicio. A bajas temperaturas (región A) la velocidad de crecimiento sigue una ley exponencial, $v \_ e^{-E_a / KT}$ 1 (E_a -1.5eV). A más altas temperaturas (región B) la velocidad de crecimiento es esencialmente independiente de la temperatura.

Para obtener una lámina epitaxial de gran calidad las temperaturas de crecimiento deben ser relativamente altas. Además, el crecimiento epitaxial debería hacerse a una temperatura a la que la velocidad de crecimiento sea insensible a variaciones de la temperatura. Por lo tanto, el crecimiento epitaxial en fase de vapor se produce generalmente en la región de transferencia de masa (altas temperaturas).

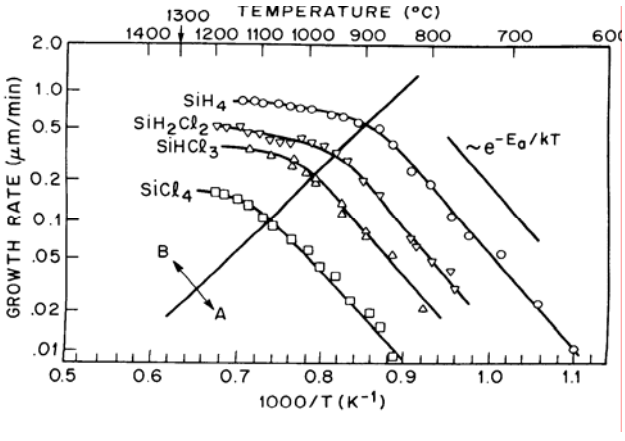


Figura 9.5.4: Velocidad de crecimiento epitaxial frente a la temperatura

9.5.3 Epitaxia por haces moleculares (M.B.E.)

La epitaxia por haces moleculares es un proceso epitaxial (crecimiento de la lámina delgada de semiconductor monocristalino) que involucra la reacción de uno o más haces térmicos de átomos o moléculas con una superficie cristalina en condiciones de alto vacío (10^{-10} Torr). De esta forma se consigue un control muy preciso tanto en la composición química cómo en los perfiles del dopante. La epitaxia por haces moleculares (MBE) tiene un gran número de ventajas comparada con la epitaxia en fase de vapor (VPE), alguna de las cuales se enumeran a continuación:

-Procesado a bajas temperaturas (400 a 800 C)

-Control preciso del perfil del dopado.

-Crecimiento de múltiples capas monocristalinas con espesores atómicos. (superredes).

La figura siguiente muestra esquemáticamente un sistema de crecimiento MBE para GaAs:

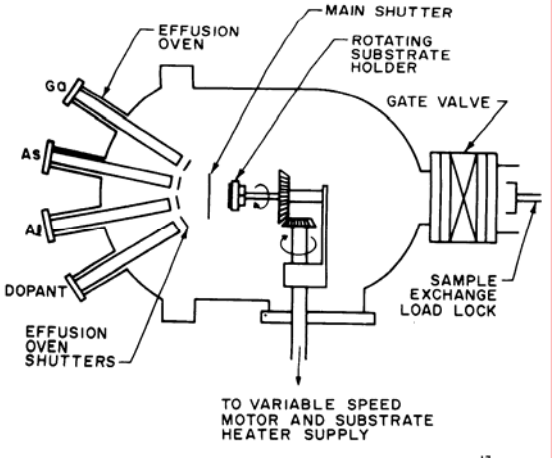


Figura 9.5.5 Esquema de un reactor MBE

Para evaporar As, Ga y los diferentes dopantes se utilizan diferentes hornos de efusión realizados con nitruro de boro pirolítico. Todos los hornos se alojan en un recipiente en el que se mantiene un alto vacío (10^{-10} Torr). La temperatura de cada horno se ajusta para dar la velocidad de evaporación deseada. El soporte del substrato gira continuamente para proporcionar capas epitaxiales uniformes.

Antes de crecer la capa epitaxial es necesario realizar una limpieza de la superficie. Esto puede hacerse "in-situ" cociendo a altas temperaturas la muestra, lo que hace que desaparezcan las impurezas por evaporación, o por difusión hacia el interior del volumen. MBE puede usar una mayor variedad de dopantes que VPE. Además, puesto que el crecimiento se hace a capas atómicas, el perfil de dopado puede ser controlado de una forma muy precisa, sin más que controlar los flujos relativos de dopantes y de silicio o arseniuro de galio.

La temperatura del substrato para el crecimiento MBE está en el rango 400 a 800 C. La velocidad de crecimiento está comprendida entre 0.001 a 0.3 μm/min.

9.6 Calidad de las capas cristalinas. Defectos, dislocaciones, impurezas residuales y otras imperfecciones.

Un cristal real, tal como una oblea semiconductora como las que acabamos de estudiar, difiere en varios aspectos de un cristal ideal:

1.- El cristal real es finito, por lo que los átomos de la superficie están incompletamente ligados, rompiéndose de esta forma la periodicidad de la red.

2.- Presenta defectos que influyen fuertemente sobre las propiedades eléctricas, mecánicas y ópticas de los semiconductores.

Hay cuatro tipos de defectos diferentes en un semiconductor: (a) puntuales ; (b) dislocaciones; (c) de área, (d) de volumen.

(a) Defectos puntuales.

Cualquier átomo extraño a la red incorporado tanto en posición substitucional (en un lugar regular) o en posición intersticial (entre lugares regulares) es un defecto puntual. Un átomo ausente de la red provoca una vacante que también es considerada como un defecto puntual. Un átomo situado en posición intersticial al lado de una vacante se denomina defecto Frenkel. Los defectos puntuales son particularmente importantes en los procesos de difusión y oxidación. La difusión de muchas impurezas depende de la concentración de vacantes, al igual que la velocidad de oxidación del silicio. Para ser eléctricamente activos, los átomos en la red deben estar colocados, por lo general, en posición substitucional, creando de esta forma niveles de energía en la banda prohibida del semiconductor.

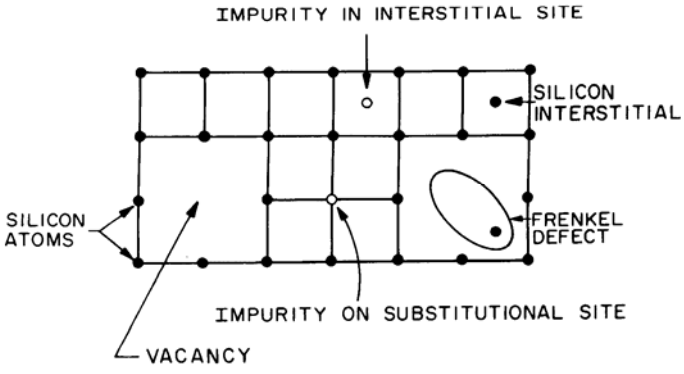


Figura 9.6.1: Tipo de defectos puntuales

(b) Dislocaciones o defectos de línea.

La figura siguiente muestra una red cristalina que presenta este tipo de defecto, que consiste en la existencia de un plano extra de átomos AB intercalado en la red. Este tipo de defectos son no-deseados puesto que actúan como lugares de precipitación para impurezas metálicas que degradan las características de los dispositivos.

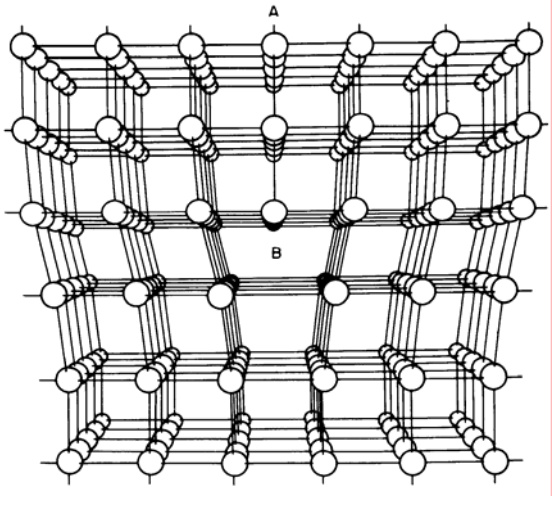


Figura 9.6.2: Ejemplo de dislocaciones en una red

(c) Defectos de área.

Los defectos de área representan una superficie de discontinuidad en la red. Existen dos defectos típicos:

-**Twins.** Representa un cambio en la orientación cristalográfica en un plano.

-**Contorno granular** aparece cuando se produce una transición entre cristales que no tienen una orientación cristalográfica particular a cristales con una orientación determinada.

Estos tipos de defectos se inducen durante el crecimiento del cristal. Los cristales que presentan alguna de estos tipos de defectos no son útiles para la construcción de circuitos integrados.

(d) Defectos de volumen.

Los precipitados de impurezas o átomos de dopantes constituyen la cuarta categoría de defectos en cristales semiconductores. Estos defectos se producen como consecuencia de la solubilidad de las impurezas en el material anfitrión. Hay una concentración específica de impurezas que la red anfitriona puede aceptar en una solución sólida del material anfitrión y la impureza. La solubilidad de la mayoría de las impurezas disminuye a medida que disminuye la temperatura. Por lo tanto, si a una temperatura dada, se introduce la concentración máxima permitida por su solubilidad de una impureza en un cristal, y a continuación este se enfría a temperaturas más bajas, la única posibilidad de alcanzar el estado de equilibrio es precipitando el exceso de átomos de impureza. Sin embargo, la diferencia de volumen entre el cristal y el precipitado de la impureza da lugar a dislocaciones.

-Impurezas residuales.

Tanto el carbono como el oxígeno son impurezas residuales (no intencionadamente introducidas) que aparecen al crecer el cristal. Las concentraciones de carbón y oxígeno son mucho mayores en los cristales crecidos mediante la técnica de Czochralski que mediante la técnica del "float zone", debido a la disolución del crisol de sílice (que proporciona oxígeno) y a la contaminación del fundido con carbono procedente del soporte de grafito.

Las concentraciones típicas de carbono están en el rango 10^{16} a 10^{17} átomos/cm^3. Carbono entra en posición substitucional en el silicio dando lugar a la formación de defectos.

En cuanto al oxígeno, las concentraciones típicas caen dentro del rango 10^{17} a 10^{18} átomos/cm^3. Al contrario que el carbono, el oxígeno tiene tanto efectos perjudiciales como beneficiosos.

-Perjudiciales. Puede actuar como donador, perturbando de esta forma la resistividad de la red proporcionada por el dopado intencionadamente introducido en el cristal.

-Beneficiosos. Por otro lado, los precipitados de oxígeno pueden ser usados para **gettering**. Por gettering se entiende el proceso consistente en eliminar impurezas no deseadas o defectos de la región de la oblea dedicada a fabricar dispositivos electrónicos. Cuando la oblea se somete a altas temperaturas (1050 C en N_2) el oxígeno se evapora de la superficie. Esto hace disminuir el contenido de oxígeno cerca de la superficie, y de esta forma la precipitación de oxígeno no ocurre cerca de la superficie de la oblea. El tratamiento crea una zona libre de defectos (zona desnuda) para la fabricación de dispositivos. Ciclos térmicos adicionales pueden utilizarse para fomentar la formación de precipitados de oxígeno en el interior de la oblea que actúan como sumideros (que atraen) impurezas, de manera que los

átomos de impurezas son atraídos hacia estos zonas, dejando el resto de las zonas, donde se fabricarán los dispositivos libres de impurezas.

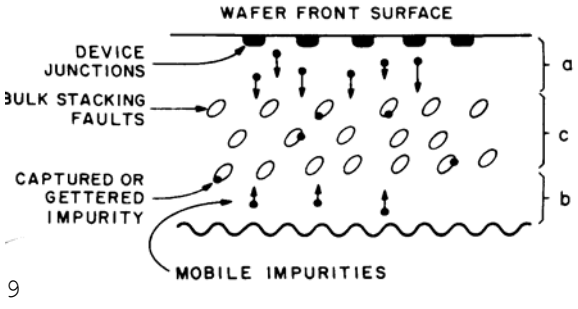


Figura 9.6.3: Gettering

La profundidad de la zona desnuda (libre de impurezas) depende del tiempo y de la temperatura de los ciclos de temperatura. La figura siguiente muestra la anchura de la zona desnuda en función del tiempo, para diferentes condiciones ambientales a las que se realiza el proceso:

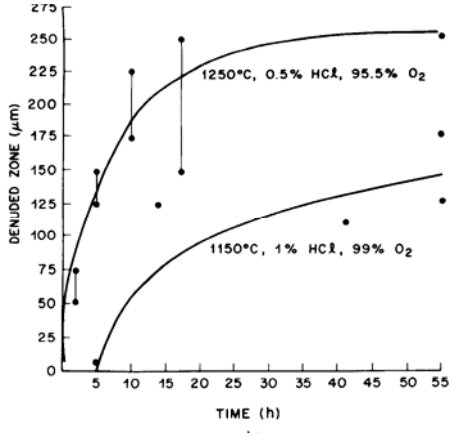


Figura 9.6.4 Espesor de la zona libre de impurezas (zona desnuda) frente al tiempo de tratamiento térmico

En cuanto al arseniuro de galio, suele estar altamente contaminado por el crisol. Sin embargo, para aplicaciones fotónicas no supone restricción alguna debido al alto valor del dopado requerido en la mayoría de los casos. Para aplicaciones de circuitos integrados, el arseniuro de galio se dopa con cromo para obtener una resistividad inicial de 10^9 Ω-cm. Oxígeno es una impureza no deseada en el arseniuro de galio debido a la formación de complejos que aumentan la resistividad de la oblea. Para minimizar la contaminación de oxígeno puede emplearse crisoles de grafito. Mediante la técnica de Bridgman se obtiene una densidad de dislocaciones un orden de magnitud más pequeña que la obtenida mediante el método de Czochralski

REFERENCIAS

[1] S. Sze. *VLSI Technology*, Ed. Mcgraw-Hill.

[2] J.D.Plummer, M.D.Deal, P.B.Griffin, *Silicon VLSI Technology*, Ed.Prentice Hall.

[3] Chang and S. Sze, *ULSI Technology*, Ed. Mcgraw-Hill.

[4] Streetman and Banerjee, *Solid State Electronic Devices*, Prentice Hall,Fifth Edition, 2000.

[5] Glosario: http://semiconductorglossary.com/

10

Capítulo

Impurificación controlada de semiconductores.

Implantador

Índice

Objetivos

- Conocer las razones por las que es necesario contaminar determinadas zonas de una muestra semiconductora.
- Describir los diferentes procedimientos existentes para la impurificación controlada de semiconductores. Establecer las principales ventajas, inconvenientes y diferencias de los diferentes procedimientos.
- Obtener modelos analíticos que nos proporcionen el perfil de dopado en función de las variables tecnológicas de proceso.
- Obtener la ecuación de la difusión de Fick y obtener soluciones a la misma una vez establecidas las condiciones iniciales y de contorno.
- Obtención de máscaras para la difusión y la implantación iónica a partir de dióxido de silicio.
- Difusión lateral.

Palabras Clave

Difusión.
Implantación Iónica.
Impurezas.
Substitucional.
Vacantes.
Intersticios.
Ecuación de Fick.
Perfil Impurezas.
Coeficiente de difusión.
Difusión Intrínseca.
Difusión extrínseca.
Difusión con concentración de dopante constante en la superficie.
Difusión con cantidad de dopante constante.
Máscaras.
Difusión lateral.
Rango.
Rango proyectado.
Parada nuclear.
Parada electrónica.
Acanalamiento.
Implantación lateral.

10.1 Introducción. Difusión e Implantación Iónica

Dopantes
Defectos puntuales, que consisten en átomos de impurezas que sustituyen a un átomo de la red cristalina, ocupando su misma posición, y que suelen tener un electrón más (donadores) o un electrón menos (aceptores), modificando de esta forma la conductividad intrínseca del cristal al añadir portadores de carga: electrones en el caso de impurezas donadoras, y huecos en el caso de impurezas aceptadoras.

Para la construcción de dispositivos electrónicos es necesario introducir localmente en el semiconductor cantidades controladas de impurezas en posición substitucional (dopantes). Para ello se utilizan generalmente alguna de las dos siguientes técnicas: **difusión** o **implantación iónica**. Con estas técnicas se puede dopar selectivamente el substrato del semiconductor y producir regiones tipo p ó tipo n según convenga.

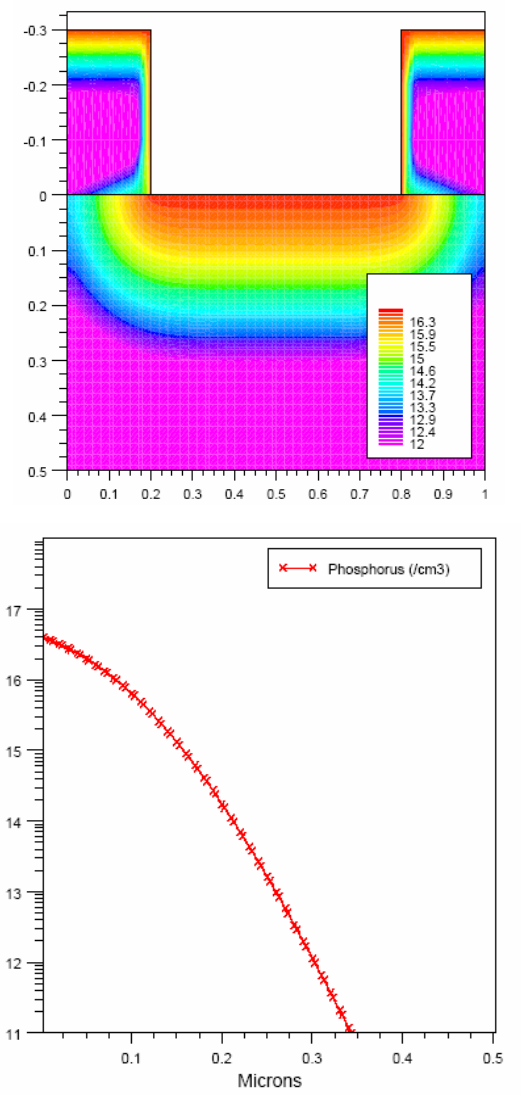
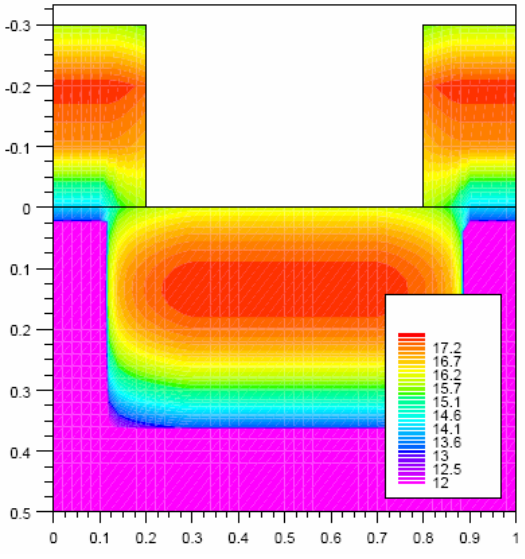


Figura 1.1 Perfil de impurezas obtenido mediante la difusión de fósforo en Silicio a 1000ºC durante 30 minutos.

Hasta comienzos de 1970, el dopado selectivo se realizaba principalmente mediante difusión a altas temperaturas. En este método los átomos de dopante se colocan en la superficie del semiconductor o cerca de ella por deposición a partir del propio dopante en fase gaseosa o bien a partir de óxidos dopados. La concentración de dopante disminuye monótonamente a medida que se aleja de la superficie, y el perfil de dopado depende tanto de la temperatura como del tiempo de difusión (Figura 10.1.1).



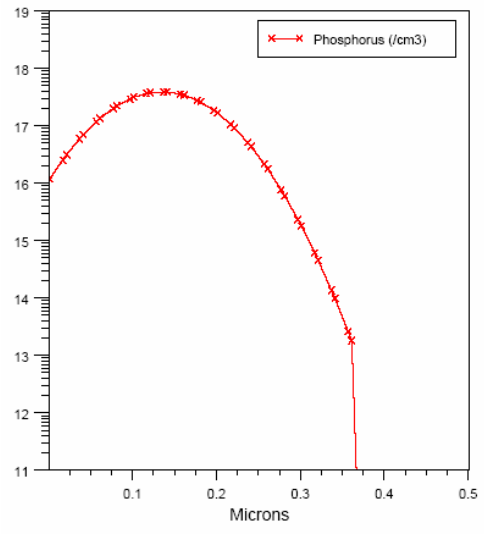


Figura 10.1.2. Perfil de impurezas resultante de la implantación iónica de fósforo en silicio.

A partir de 1970, muchas operaciones de dopado selectivo han sido realizadas mediante técnicas de implantación iónica. En esta técnica, los átomos del dopante son implantados en el interior del semiconductor por medio de haces iónicos de alta energía. El perfil del dopado tiene un máximo en el interior del semiconductor (Figura 10.1.2), y está determinado por la masa de los iones y la energía con que se hacen incidir los mismos sobre la superficie semiconductora. Las ventajas de la implantación iónica sobre la difusión son un control preciso de la cantidad de dopantes introducidos, la reproductibilidad de los perfiles de impurezas y una menor temperatura de proceso.

10.2 Difusión. Ecuación de Fick

Para producir la difusión de impurezas en el interior del semiconductor, se colocan las obleas del mismo en el interior de un horno a través del cual se hace pasar un gas inerte portador que contenga el dopante deseado. El sistema es similar al empleado en la oxidación térmica. Los rangos de temperatura van entre 800 y 1200ºC para el silicio y 600 y 1000ºC para el arseniuro de galio. (Figura 10.2.3)

Para el silicio, el dopante más usual es el boro en el caso de impurezas tipo p, y el arsénico y fósforo en el caso de impurezas tipo n. Estos tres elementos tienen una alta solubilidad en silicio ($5x10^{20}$ cm^{-3}) en el rango de temperatura de difusión, como muestra la figura 10.2.4. La introducción de estos dopantes puede hacerse de muy diferentes formas a partir de fuentes sólidas (BN, As_2O_3, P_2O_5), líquidas (BBr_3, $AsCl_3$, $POCl_3$) y gaseosas (B_2H_6, AsH_3 y PH_3). Generalmente el material elegido es transportado hasta la superficie del semiconductor por un gas inerte (N_2).

Para la difusión de impurezas en el arseniuro de galio, hay que utilizar técnicas especiales debido a la alta presión de vapor del arsénico, para evitar la pérdida de éste por evaporación. Estos métodos especiales incluyen la difusión en ampollas selladas con una alta sobrepresión de arsénico, y la difusión en un horno abierto con una capa protectora de óxido dopado. Para dopado tipo p se utilizan aleaciones de Zn-Ga-As y ZnAs2 en el caso de ampollas selladas y ZnO-SiO2 para el horno abierto. Como dopantes tipo n se utilizan tanto el azufre como el selenio. La figura 10.2.5 muestra una fotografía de un horno de difusión.

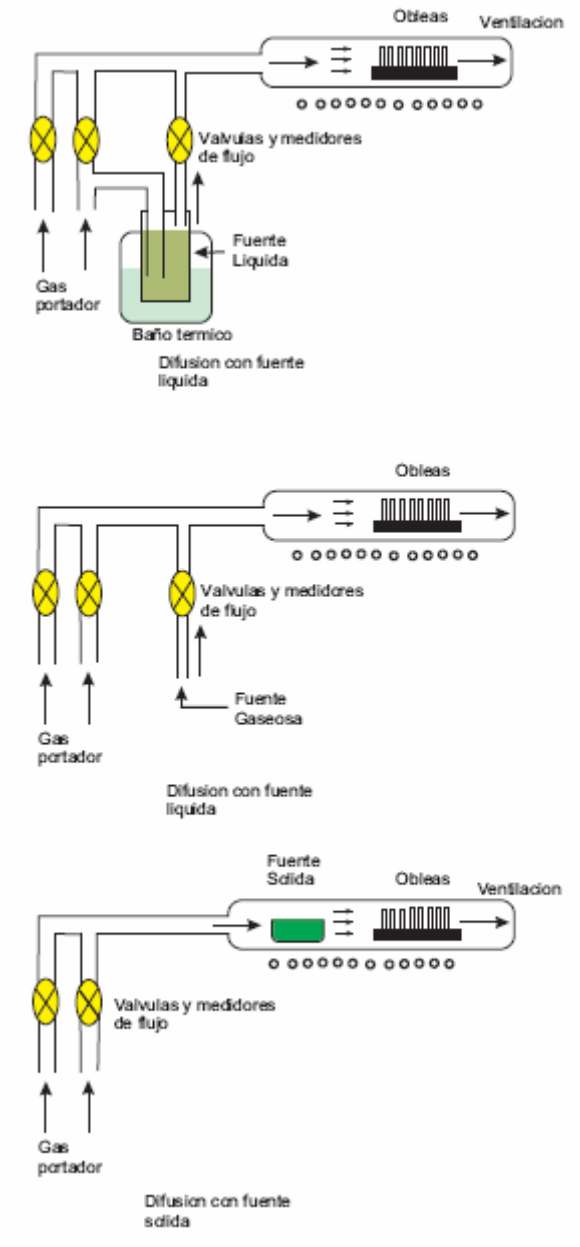


Figura 10.2.3.- Diferentes sistemas de difusión, con fuentes líquidas, gaseosas y sólidas.

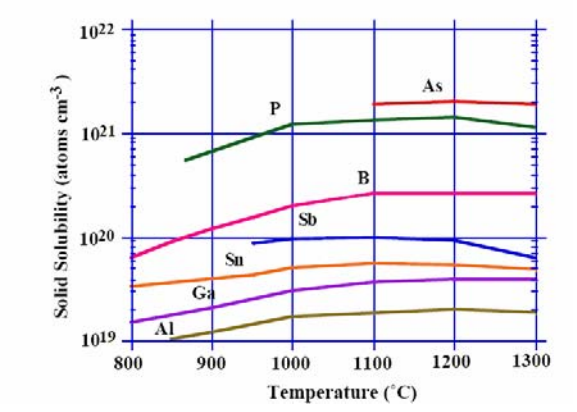


Figura 10.2.4. Solubilidad de diferentes sustancias en silicio en función de la temperatura.

Figura 10.2.5. Horno de difusión con el módulo de control a la izquierda y cuatro reactores de difusión.

Ecuación de Fick.

El proceso de difusión en un semiconductor puede verse como el movimiento de los átomos de dopantes por la red cristalina a través de las vacantes e intersticios de la misma, como consecuencia de un gradiente de concentración de impurezas, como muestra la figura 10.2.6

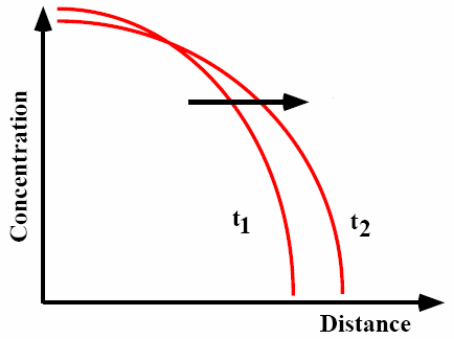
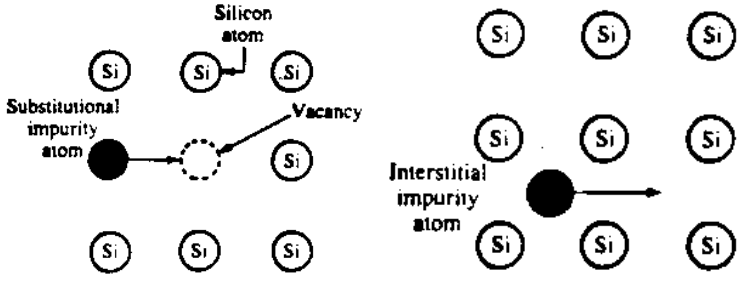


Figura 10.2.6. La difusión de las impurezas es provocada por un gradiente de concentración.

A temperaturas elevadas los átomos de la red vibran en torno a su posición de equilibrio. Existe por lo tanto una probabilidad finita de que el átomo que forma la red adquiera la suficiente energía para abandonar su lugar en la red y pase a ocupar un intersticio dejando una vacante. Cuando un átomo de impurezas ocupa la vacante dejada por el átomo anfitrión, el proceso se denomina **difusión por vacante** (a). Si un átomo intersticial se mueve de un lugar a otro del cristal sin ocupar un sitio de la red, entonces el mecanismo se denomina **difusión por intersticios (b).** La figura 10.2.7 muestra los dos mecanismos de difusión:



(a) (b)

Figura 2.7 Mecanismos de difusión: (a) por vacantes; (b) por intersticios.

En general ambos mecanismos de difusión pueden estar presentes en el movimiento de las impurezas a través de un substrato cristalino, aunque dependiendo del tipo de impureza o substrato puede dominar uno u otro mecanismo. En general, la facilidad de una impureza

a moverse a través de un substrato se caracteriza por el coeficiente de difusión D.

Para obtener una ecuación que nos relacione la evolución de la concentración de impurezas con el tiempo como consecuencia del movimiento de impurezas en el interior del semiconductor, consideremos la figura 10.2.8. Si C(x) es la concentración de impurezas en un punto x del cristal, supondremos que el flujo de impurezas en ese punto x, F(x), que atraviesa una superficie unidad por unidad de tiempo es proporcional a C(x):

$$F(x) = -D\frac{\partial C(x)}{\partial x} \tag{10.1}$$

donde la constante de proporcionalidad D se denomina coeficiente de difusión o difusividad. De acuerdo con esto, el flujo de átomos de dopante es proporcional al gradiente de concentración, y éstos se moverán desde regiones con alta concentración a regiones con más baja concentración, tal como indica el signo menos.

Consideremos una capa infinitesimal de semiconductor de espesor dx como muestra la figura 10.2.8

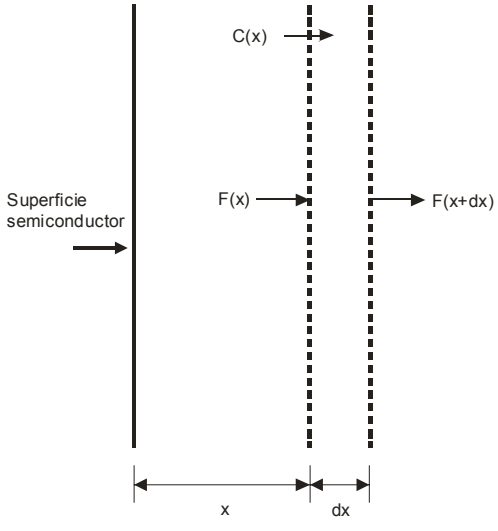


Figura 10.2.8. Modelo de Fick para la difusión de impurezas.

El cambio en la concentración C(x) de átomos de dopante con el tiempo en dx a una distancia x de la superficie puede escribirse como la diferencia de flujo de átomos de dopante que entran en dicha región por la izquierda y el flujo de átomos de dopante que salen de ella por la derecha

$$\frac{\partial C(x,t)}{\partial t}dx = F(x) - F(x+dx) \qquad (10.2)$$

Desarrollando en serie de Taylor, F(x+dx) hasta primer orden y sustituyendo se obtiene:

$$\frac{\partial C(x,t)}{\partial t} = -\frac{\partial F(x)}{\partial x} = \frac{\partial}{\partial x}\left(D\frac{\partial C(x,t)}{\partial x}\right) \qquad (10.3)$$

Si el coeficiente de difusión, D, es constante, se obtiene finalmente:

$$\frac{\partial C(x,t)}{\partial t} = D\frac{\partial^2 C(x,t)}{\partial x^2} \qquad (10.4)$$

Esta ecuación se conoce con el nombre de **ecuación de difusión de Fick,** y puede emplearse para calcular la concentración de dopante en todo instante en el interior del cristal, una vez que se fijen las condiciones iniciales y de contorno para poder resolverla.

La figura 10.2.9 muestra los coeficientes de difusión empíricos, medidos para bajas concentraciones en silicio y arseniuro de galio. Se observa una dependencia del tipo exponencial:

$$D = D_0 e^{-\frac{E_a}{kT}} \qquad (10.5)$$

donde D_0 es el coeficiente de difusión para temperatura infinita y E_a la energía de activación. Para el modelo de difusión por intersticios, E_a está relacionada con la energía requerida para mover los átomos de dopante de un intersticio a otro; en este caso, los valores de E_a caen entre 0.5 y 1.5eV, tanto para el silicio como para el arseniuro de galio. Por el contrario, para el modelo de difusión por vacantes, la energía de activación, E_a está relacionada tanto con la energía de movimiento como con la energía de formación de vacantes. Por ello, para la difusión por vacantes, la energía de activación es mayor que en el caso de difusión por intersticios, tomando valores entre 3 y 5eV. En la figura 10.2.9, puede observarse que para los dopantes con mayor coeficiente de difusión (parte superior de las gráficas (cobre, litio, sodio)) la pendiente de las curvas, esto es, la energía de activación, E_a, es menor de 2eV, siendo el movimiento atómico intersticial el mecanismo de difusión dominante. Sin embargo, para las impurezas con menor coeficiente de difusión (parte inferior de las

gráficas) la energía de activación es mayor de 3eV, siendo la difusión por vacantes el mecanismo principal.

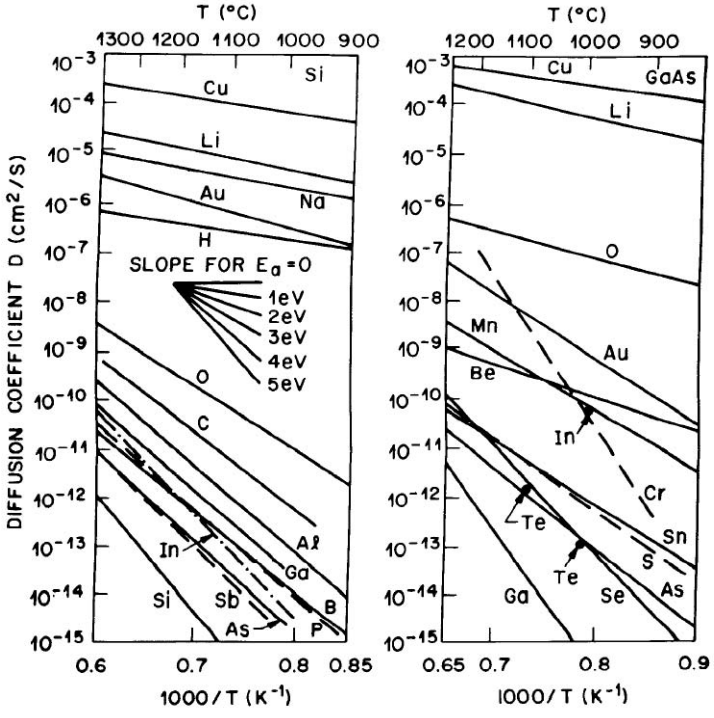


Figura 10.2.9. Coeficientes de difusión empíricos de diferentes impurezas en silicio y arseniuro de galio frente al inverso de la temperatura.

10.3 Perfiles de difusión

El perfil del dopado obtenido a través de difusión depende de las condiciones iniciales y de contorno. Vamos a estudiar en este apartado dos casos importantes de difusión:

1.-Difusión con concentración de dopante constante en la superficie.

Los átomos de impureza son transportados a partir de una fuente gaseosa hacia la superficie semiconductora. La fuente gaseosa mantiene a nivel constante la concentración de superficial durante todo el proceso de difusión.

2.- Difusión con cantidad total de dopante constante.

Se deposita inicialmente una cantidad fija de dopante sobre la superficie de la oblea. Este dopante se difunde a continuación hacia el interior de la oblea. La concentración en la superficie disminuye a medida que transcurre el proceso de difusión.

Difusión con concentración de dopante constante en la superficie

Para poder resolver la ecuación de Fick , (Eq.10.4) necesitamos fijar condiciones iniciales y las condiciones de contorno.

Bajo la suposición bajo la cual se obtuvo la ecuación 10.4 de que el coeficiente de difusión no depende de la concentración, sin pérdida de generalidad podemos fijar como condición inicial:

$$C(x, t = 0) = 0 \tag{10.6}$$

esto es, inicialmente la concentración de dopante en el semiconductor es cero. En cuanto a las condiciones de contorno:

$$C(x = 0, t) = cte = C_S \tag{10.7}$$

$$C(x = \infty, t) = 0 \tag{10.8}$$

donde C_s es la concentración en la superficie (x=0) que es independiente del tiempo (esta condición es la que da nombre al método). La segunda condición de contorno establece que a grandes distancias de la superficie no hay átomos de dopante.

La solución a la ecuación de difusión de Fick con estas condiciones iniciales y de contorno viene dada por:

$$C(x, t) = C_S erfc\left(\frac{x}{2\sqrt{Dt}}\right) \tag{10.9}$$

donde **erfc** es la función error complementaria y $\sqrt{Dt}$ la longitud de difusión. La figura siguiente muestra, tanto en escala lineal como en escala logarítmica la concentración normalizada de impurezas como función de la profundidad para tres valores diferentes de la longitud de difusión (D se mantiene fija). Se observa que a medida que aumenta el tiempo el dopante penetra cada vez más hacia el interior del semiconductor.

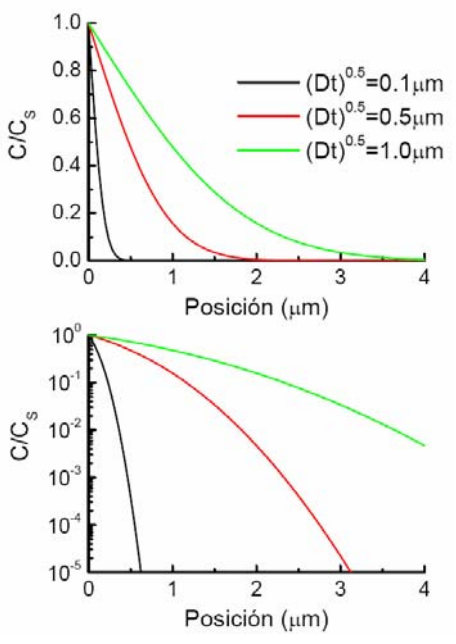


Figura 10.3.1. Perfiles de difusión con concentración de dopante constante para diferentes tiempos de difusión.

El número total de átomos de dopante por unidad de área en el semiconductor viene dado al integrar la concentración de dopante a toda la estructura:

$$Q(t) = \int_0^\infty C(x,t)dx \tag{10.10}$$

Teniendo en cuenta la forma de C(x,t) se obtiene finalmente:

$$Q(t) = \frac{2}{\sqrt{\pi}} C_S \sqrt{Dt} \approx 1.13\sqrt{Dt} \tag{10.11}$$

La ecuación 10.11 pone de manifiesto que la cantidad de dopante aumenta a medida que aumenta el tiempo. Finalmente el gradiente del perfil de difusión puede obtenerse derivando la concentración de átomos de dopante (Ecuación 10.9) teniendo en cuenta las propiedades de la función error complementaria:

$$\left.\frac{dC}{dx}\right|_{x,t} = -\frac{C_S}{\sqrt{\pi Dt}} e^{-\frac{x^2}{4Dt}} \tag{10.12}$$

Difusión con cantidad total de dopante constante

En este caso se deposita una cantidad fija de dopante sobre la superficie del semiconductor que va difundiéndose por el mismo poco a

poco. La condición inicial de este problema es la misma que en el caso anterior, es decir, suponemos que no hay inicialmente átomos de dopante en el interior del semiconductor:

$$C(x, t = 0) = 0 \tag{10.13}$$

Las condiciones de contorno por el contrario son ahora:

$$\int_0^\infty C(x,t)dx = S \tag{10.14}$$

$$C(x = \infty, t) = 0 \tag{10.15}$$

donde S es la cantidad total de dopante por unidad de área. Con estas condiciones de contorno, la solución a la ecuación de Fick (Ecuación 10.4) es una distribución gaussiana:

$$C(x,t) = \frac{S}{\sqrt{\pi Dt}} e^{-\frac{x^2}{4Dt}} \tag{10.16}$$

Puesto que la cantidad total de dopante permanece constante, a medida que este se difunde hacia el interior del semiconductor deberá disminuir su concentración en la superficie:

$$C(x = 0, t) \equiv C_S(t) = \frac{S}{\sqrt{\pi Dt}} \tag{10.17}$$

La figura 10.3.2 muestra el perfil de la distribución normalizada por la cantidad total de dopante. Se observa en la figura la reducción de la concentración superficial a medida que aumenta el tiempo de difusión, y se ensancha el perfil de la distribución. El gradiente del perfil del dopado viene dado por:

$$\left.\frac{dC}{dx}\right|_{x,t} = -\frac{xS}{2\pi(Dt)^{3/2}} e^{-\frac{x^2}{4Dt}} \tag{10.18}$$

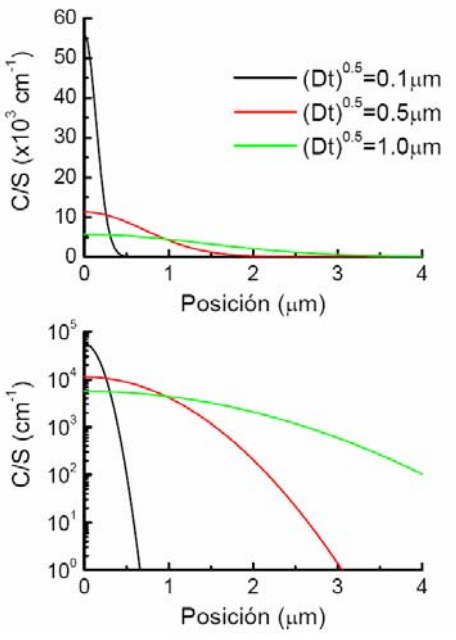


Figura 10.3.2. Perfiles de difusión con cantidad total de dopante constante para diferentes tiempos de difusión.

Las dos tipos de difusiones que hemos estudiado (función error y función gaussiana) son funciones de la distancia normalizada $x / 2\sqrt{Dt}$. Por lo tanto si se normaliza la concentración de dopado con la concentración en la superficie, podemos representar cada distribución con una única curva válida para todo t. Hay que hacer notar que mientras para la función error complementaria C_s es una constante, para la distribución gaussiana, la variable de normalización depende del tiempo (Figura 10.3.3).

En la fabricación de circuitos integrados, generalmente se utiliza un proceso de difusión de dos pasos: En un primer paso se lleva a cabo un proceso de difusión con concentración superficial constante (**predeposición).** A continuación se lleva a cabo una difusión con cantidad de dopante constante (**redistribución**). Para la mayoría de los casos prácticos, la longitud de difusión para la etapa de predeposición es mucho más pequeña que la longitud de difusión para la etapa de distribución. Por lo tanto, podemos considerar que el perfil de impurezas conseguido mediante la etapa de predeposición es una función delta en la superficie.

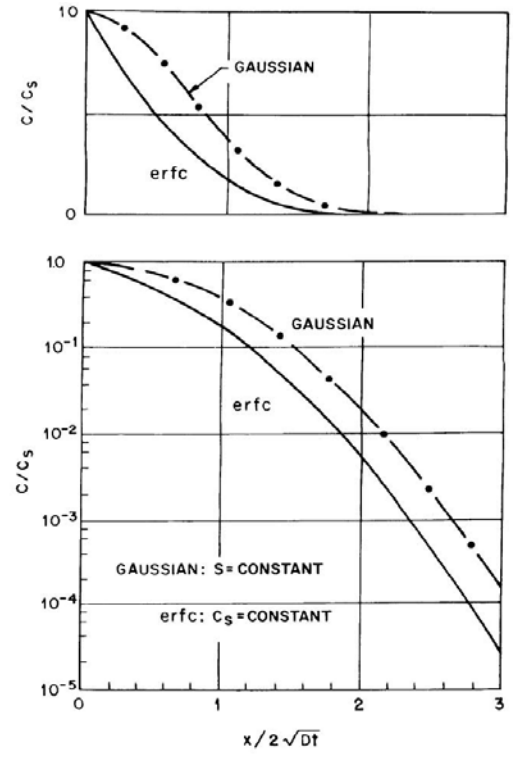


Figura 10.3.3. Concentraciones normalizadas frente a distancias normalizadas para la función error complementada (erfc) y la distribución gaussiana.

10.4 Difusión extrínseca

Hasta ahora hemos considerado que el coeficiente de difusión es constante. Esto ocurre únicamente cuando la concentración de dopado es más pequeña que la concentración intrínseca del semiconductor, n_i a la temperatura de difusión. Por ejemplo para T=1000ºC, $n_i=5x10^{18}cm^{-3}$ para el GaAs y $n_i=5x10^{17}cm^{-3}$ para Si. Cuando la concentración de dopado es más pequeña que la concentración intrínseca se habla entonces de **difusión intrínseca**, para la cual los coeficientes de difusión son constantes y se obtienen los perfiles que acabamos de estudiar. Por el contrario si la concentración del dopado es mayor que la concentración intrínseca n_i a la temperatura de difusión, entonces se habla de **difusión extrínseca**, y los coeficientes de difusión dependen de la concentración de impurezas, tal como muestra la figura 10.4.1, esto es, produciendose un aumento del coeficiente de difusión al aumentar la concentración de dopantes.

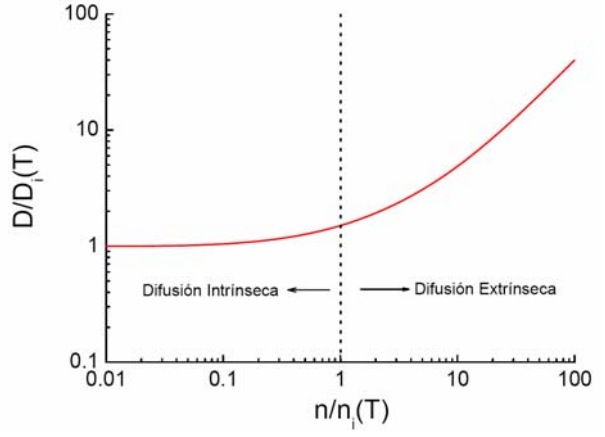


Figura 10.4.1. Evolución del coeficiente de difusión en función de la concentración de dopantes.

En la región intrínseca, los perfiles de dopado de difusiones secuenciales o simultáneas, pueden determinarse por superposición, es decir, las difusiones pueden tratarse independientemente.

Cuando los coeficientes de difusión dependan de la concentración de impurezas, no podremos utilizar la ecuación de difusión de Fick para determinar el perfil del dopado, y habrá que utilizar por contra la ecuación de Fick generalizada:

$$\frac{\partial C(x,t)}{\partial t} = \frac{\partial}{\partial x}\left(D \frac{\partial C(x,t)}{\partial x} \right) \tag{10.19}$$

Supongamos el caso en el que el coeficiente de difusión pueda escribirse en la forma:

$$D = D_S \left(\frac{C}{C_S} \right)^{\gamma} \tag{10.20}$$

donde C_S es la concentración en la superficie, D_S es el coeficiente de difusión en la superficie y γ un entero positivo. En este caso, la ecuación de difusión proporciona una ecuación diferencial ordinaria que puede resolverse numéricamente. La figura 10.4.2 proporciona las soluciones en condiciones de concentración constante en la superficie para diferentes valores de γ.

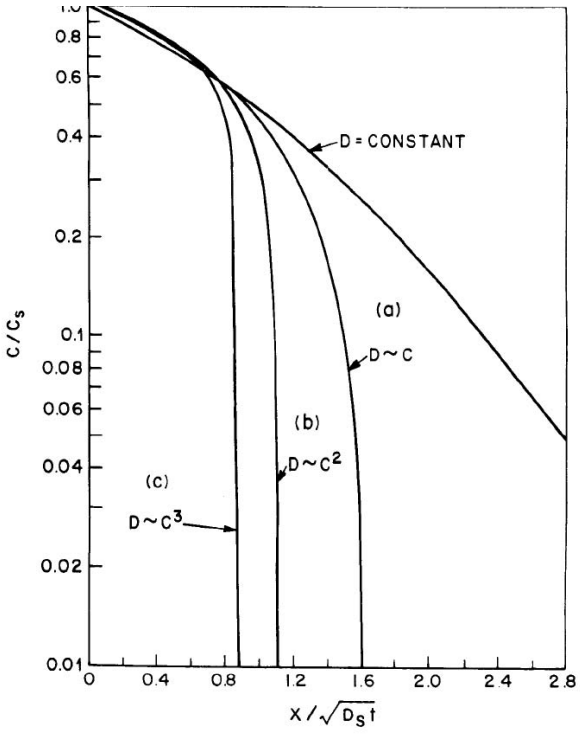


Figura 10.4.2. Diferentes perfiles normalizados de difusión extrínseca.

En la figura 10.4.2 se aprecia que para cuando el coeficiente de difusión depende de la concentración, los perfiles de difusión son mucho más abruptos que los obtenidos cuando el coeficiente de difusión es constante.

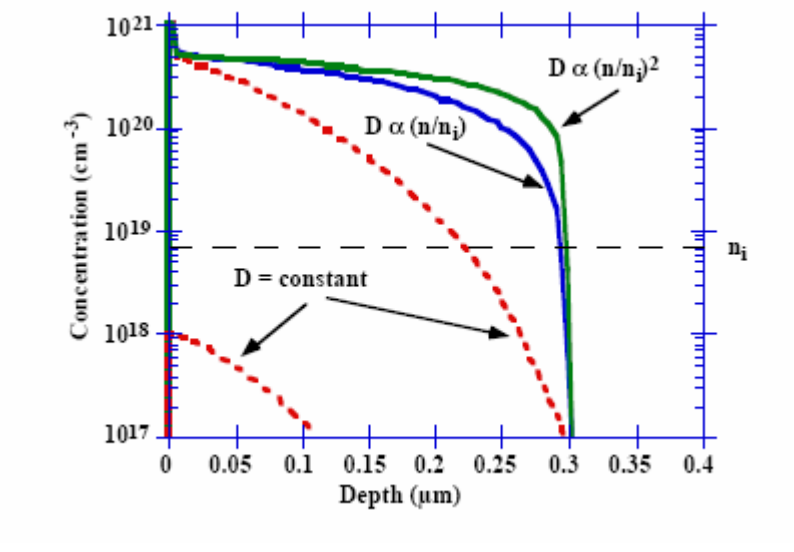


Figura 10.4.3. Perfiles de difusión para una difusión extrínseca.

Los coeficientes de difusión de las impurezas más usadas en silicio dependen con la concentración: arsénico y boro depende linealmente con la concentración, mientras que el coeficiente de difusión del fósforo depende del cuadrado de la concentración, por lo que proporcionan uniones muy abruptas cuando se difunden sobre un dopado con impurezas del tipo contrario. El fósforo tiene una alta difusividad en silicio de ahí que se utilice para formar impurezas profundas.

10.5 Efectos relacionados con la difusión

Efecto de difusiones sucesivas

Frecuentemente, en el proceso de fabricación de un circuito integrado hay diferente procesos de difusión. Cada proceso de difusión supone el someter toda la oblea a una temperatura T durante un periodo de tiempo t. Las impurezas introducidas previamente en el semiconductor por procesos de difusión o implantación anteriores también se moverán modificando su concentración.

Si todos los procesos de difusión se producen a la misma temperatura, donde el coeficiente de difusión es constante, entonces la longitud de difusión viene dado por:

$$(Dt)_{eff} = D_1(t_1 + t_2 + \ldots) = D_1 t_1 + D_1 t_2 + \ldots \qquad (10.22)$$

La expresión 10.21 viene a decir que hacer una difusión con un tiempo total t_1+t_2 es lo mismo que hacer una difusión durante un tiempo t_1 y posteriormente otra difusión, a la misma temperatura durante un tiempo t_2.

Matemáticamente, podríamos aumentar el tiempo t_2 en la expresión 10.21 multiplicando por un factor D_2/D_1, quedando:

$$(Dt)_{eff} = D_1 t_1 + D_1 t_2 \left(\frac{D_2}{D_1}\right) = D_1 t_1 + D_2 t_2 \qquad (10.22)$$

y obtener así una expresión para la longitud de difusión efectiva total para un dopante que se difunde a una temperatura T_1 con difusividad D_1 durante un tiempo t_1 y posteriormente se difunde a una temperatura T_2 con una constante de difusión D_2 durante un tiempo t_2. En general, la longitud de difusión total viene dada por la suma de las longitudes de difusión de cada proceso:

$$(Dt)_{eff} = D_1 t_1 + D_2 t_2 + \ldots = \sum_i D_i t_i \qquad (10.23)$$

Máscaras para la difusión de SiO_2

Las difusividades de los dopantes usualmente utilizados son mucho más pequeñas en SiO_2 que en silicio. Por lo tanto, el dióxido de silicio puede emplearse como una máscara efectiva de aislamiento frente al paso de impurezas. Este efecto es la base del desarrollo de la moderna tecnología electrónica: Se deposita una capa de óxido sobre la superficie del semiconductor. En este óxido se abren ventanas mediante un proceso de etching o grabado. De esta forma se consigue incorporar impurezas en áreas selectivas del dopado.

En el caso del dióxido de silicio el proceso de difusión puede describirse mediante las dos siguientes etapas. Durante la primera etapa las impurezas del dopante reaccionan con el dióxido de silicio para formar vidrio. A medida que el proceso continúa el espesor del vidrio aumenta hasta que todo el dióxido de silicio se convierta en vidrio. Una vez formado el vidrio, las impurezas de dopante se difunden por el vidrio, y tras alcanzar la interfase vidrio-silicio, penetran y se difunden por el interior del silicio. Durante la primera etapa el dióxido de silicio es totalmente efectivo en proteger el silicio de las impurezas del dopante. En consecuencia, el espesor del óxido requerido para la protección del silicio viene determinada por la velocidad de formación del vidrio la que a su vez está determinada por el coeficiente de difusión de las impurezas en el dióxido de silicio. Las constantes de difusión típicas en dióxido de silicio de los dopantes más utilizados son, a 900 ºC, $4x10^{-19}cm^2/s$ para arsénico $3x10^{-19}cm^2/s$ para boro, y $10^{-18}cm^2/s$ para fósforo, mucho más pequeñas que en silicio como puede observarse en la figura 10.2.9.

La figura 10.5.1 compara los perfiles de fósforo obtenidos mediante una difusión a T=1000ºC durante t=30 minutos sobre SiO_2 (negro) y sobre Si (rojo). Obsérvese cómo la penetración de impurezas de boro en la capa de dióxido de silicio es mucho menor que en el caso de silicio. La figura 10.5.2 muestra el espesor mínimo de óxido térmico crecido en ambiente seco requerido para formar una máscara efectiva para la difusión de boro y fósforo en función del tiempo y de la temperatura. Puesto que la constante de difusión del fósforo en el SiO_2 es mayor que la del boro, se necesitan espesores más grandes de óxido para fabricar las máscaras para la difusión del fósforo.

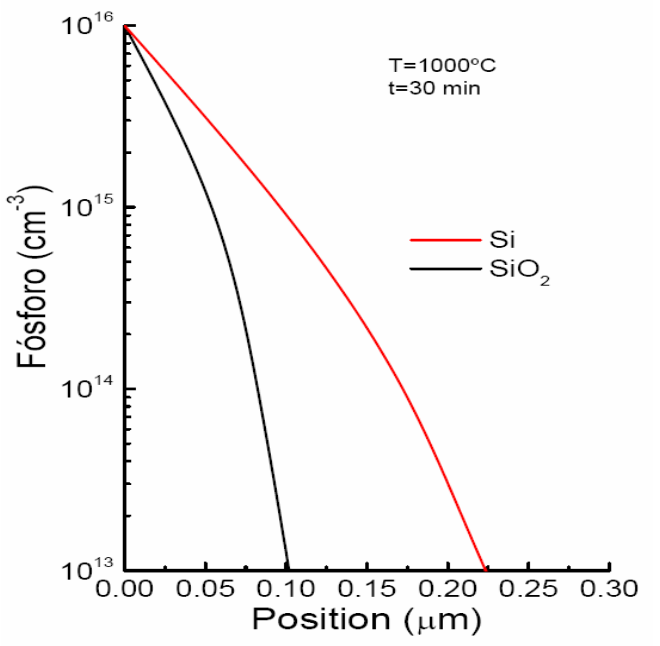


Figura 10.5.1. Pérfiles de impurezas obtenidos mediante la difusión de fósforo en Silicio y SiO_2.

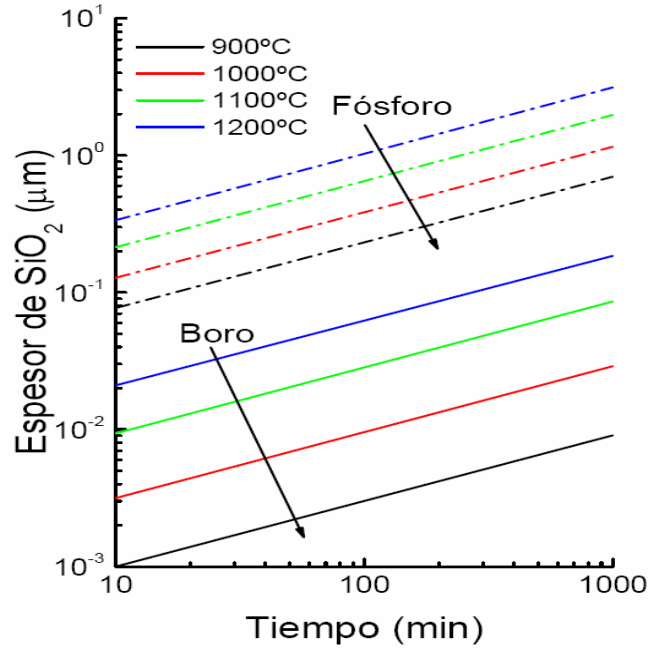


Figura 10.5.2. Espesor mínimo de la máscara de SiO_2 requerido para impedir la difusión de boro y fósforo en el substrato.

Redistribución de impurezas durante la oxidación térmica

Al producirse la oxidación térmica del silicio, se consume el silicio próximo a la superficie. Como consecuencia se produce la redistribución de las impurezas contenidas en el silicio. Esta redistribución depende de diferentes factores. Un primer factor es el conocido coeficiente de segregación. Cuando dos fase sólidas se ponen en contacto (silicio + oxido) las impurezas en ambos sólidos se redistribuyen hasta alcanzar el equilibrio. La relación entre las concentraciones de impurezas en el silicio y en el dióxido de silicio se conoce con el nombre de coeficiente de segregación, k :

$$k = \frac{\textbf{Concentración en equilibrio impurezas en Si}}{\textbf{Concentración en equilibrio impurezas en } \mathbf{SiO_2}} \qquad (10.23)$$

Si el factor de segregación para una determinada impureza es mayor que 1, k>1, las impurezas son rechazadas del óxido hacia el silicio, aumentando la concentración de impurezas en el silicio. Si k<1, entonces las impurezas se mueven rápidamente hacia el SiO_2, disminuyendo la concentración de impurezas en el silicio.

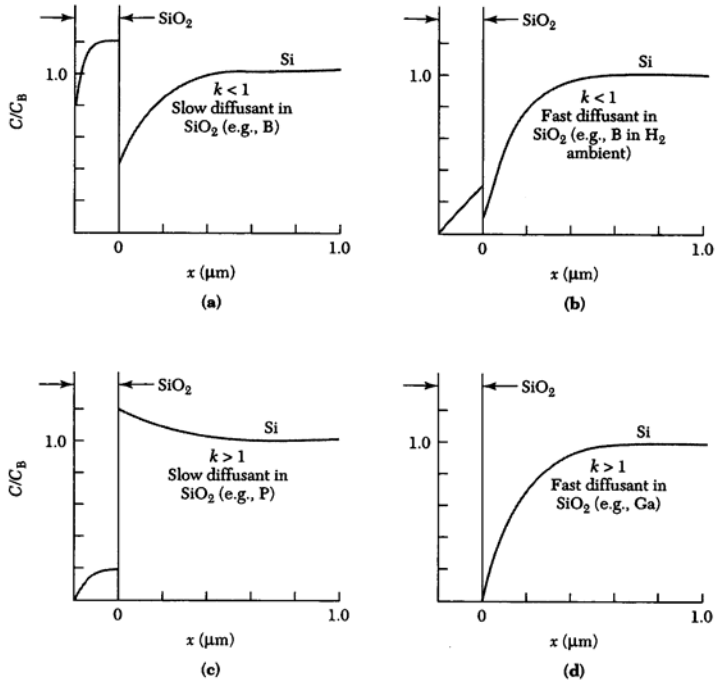


Figura 10.5.3. Redistribución de impurezas en silicio como consecuencia de la oxidación térmica.

Hay que tener en cuenta un segundo factor que influye sobre la redistribución de impurezas, y es la facilidad con la que las impurezas se puedan mover a través del SiO_2 y escapar al ambiente gaseoso. La figura 10.5.3 resume las posibles combinaciones entre los factores anteriores, mostrando los perfiles de impurezas resultantes tras la oxidación.

En el caso de silicio, los dopantes tipo N, fósforo y arsénico presentan un coeficiente de segregación mayor que 1, por lo que las impurezas son rechazadas del SiO_2 al Si, produciéndose el aumento de la concentración en la superficie del silicio. Por el contrario, el boro presenta un coeficiente de segregación menor que la unidad, y por lo tanto se produce una disminución de la concentración de impurezas en la superficie del silicio (Figura 10.5.4).

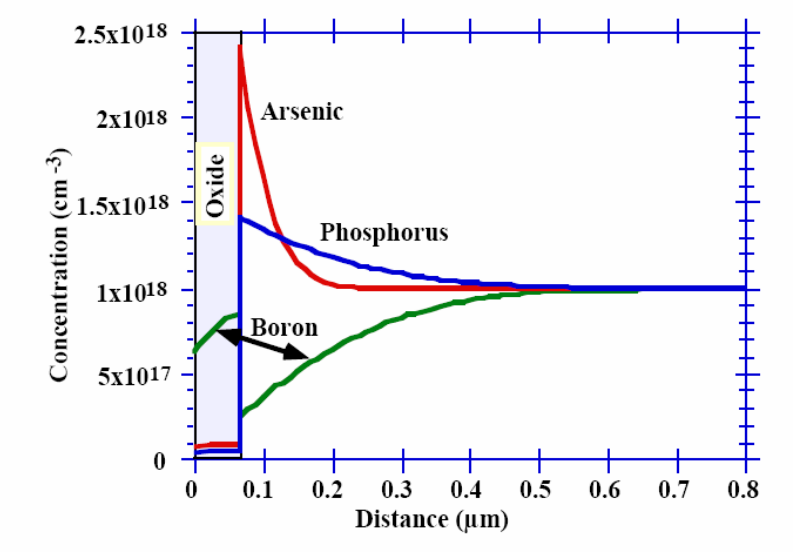


Figura 10.5.4 Perfiles de boro, fósforo y arsénico en silicio tras una oxidación térmica.

Redistribución de impurezas durante el crecimiento epitaxial.

Al igual que en el caso de la oxidación térmica, también se produce una redistribución de impurezas durante el crecimiento epitaxial. De esta forma el perfil de impurezas real se separa del perfil abrupto. Durante el crecimiento epitaxial, la temperatura es tan elevada que se produce la difusión de impurezas tanto hacia el substrato como hacia la capa epitaxial crecida, como se muestra en la figura 10.5.5 donde se representa los perfiles de boro y fósforo, cuando sobre un substrato de silicio tipo p con un dopado $N_A=5x10^{15}cm^{-3}$, se crece una capa epitaxial de silicio tipo n, con dopado $N_D=2x10^{16}cm^{-3}$.

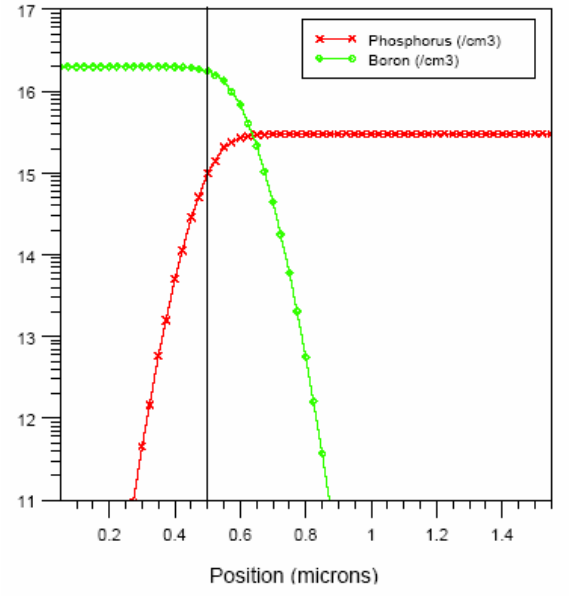


Figura 10.5.5. Perfil de boro y fósforo obtenidos tras el crecimiento epitaxial de una capa de silicio de 0.5µm de espesor.

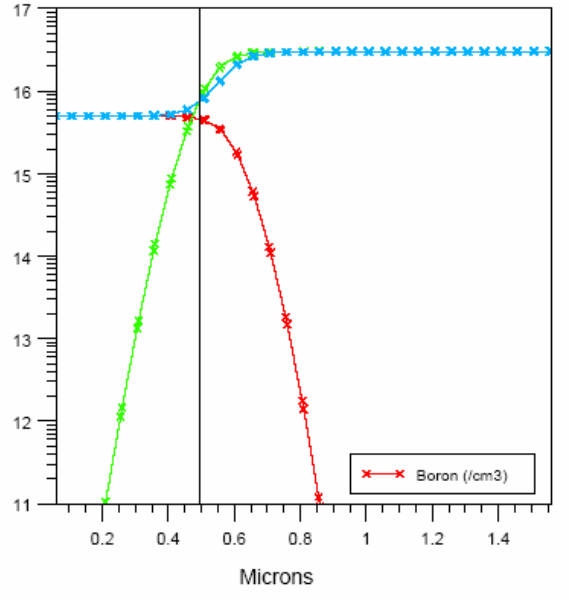


Figura 10.5.6. Perfil de boro obtenido tras el crecimiento epitaxial de una capa de silicio de 0.5µm de espesor

La figura 10.5.6 muestra, el perfil total obtenido (en azul) cuando sobre un substrato tipo p con una concentración $N_{A\text{-}sub}=3x10^{16}cm^{-3}$, se crece una capa epitaxial de silicio tipo p, pero con una concentración menor de $N_{A\text{-}epi}=5x10^{15}cm^{-3}$. Obsérvese como el perfil total se separa de

un perfil escalonado como consecuencia de la redistribución de impurezas durante el crecimiento epitaxial.

Difusión lateral

La ecuación de difusión unidimensional que hemos considerado hasta ahora no es válida en los extremos de las ventanas abiertas en el óxido de máscara. En estos puntos las impurezas se difundirán tanto hacia abajo, como lateralmente, por lo que habrá que considerar un tratamiento bidimensional. La figura siguiente muestra las líneas de concentración de dopante constante para el caso de difusión con concentración constante en la superficie, y difusividad independiente de la concentración.

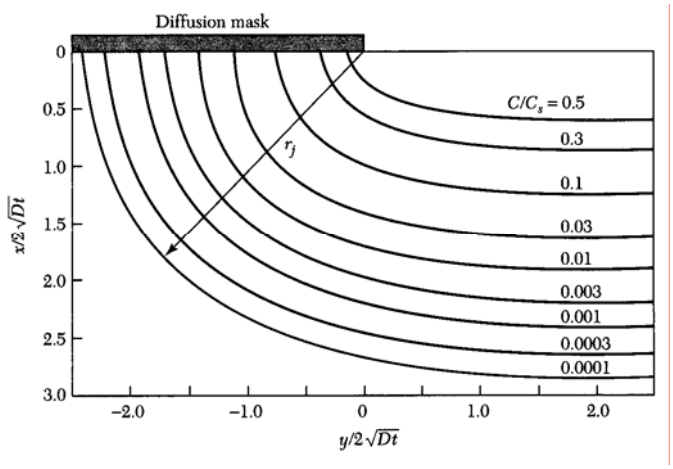


Figura 10.5.7. Contornos de difusión en torno al extremo de una ventana abierta en la máscara de óxido.

La figura 10.5.7 muestra que la penetración lateral es del orden del 80% de la profundidad alcanzada en la dirección vertical. Para difusividades dependientes de la concentración la relación entre la penetración lateral y penetración vertical es del 65 al 70 %. Debido a los efectos de la difusión lateral, las uniones (material tipo n difundido en un material tipo p o viceversa) consiste en una zona plana con extremos cilíndricos con un radio de curvatura r_j como muestra la figura anterior. (Ojo al tener en cuenta la longitud por ejemplo del canal en un MOSFET (longitud de máscara y longitud efectiva).

10.6 Implantación Iónica. Energía y dosis

La implantación iónica consiste en la introducción de partículas energéticas, cargadas, en el interior de un substrato cómo el silicio. La energía de los iones va del rango 30 a 300 keV. Las concentraciones típicas de iones implantados por centímetro cuadrado es de 10^{11} a 10^{16} iones/cm^2. Las principales ventajas de la implantación iónica sobre la

difusión es el control más preciso de los perfiles de dopado deseados y la baja temperatura a la que se realiza el proceso.

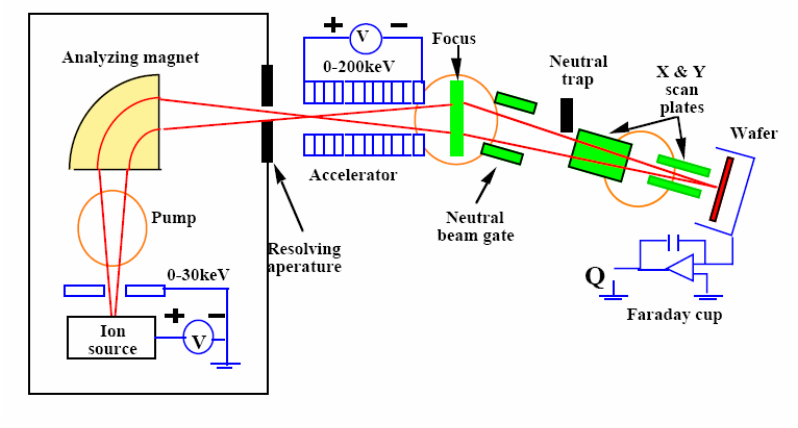


Figura 10.6.1. Esquema de un implantador iónico.

La fuente de iones contiene los átomos ionizados del dopante. Estos átomos pasan a través de un analizador magnético de masas donde los iones no deseados son eliminados del haz de iones. A continuación el haz de iones entran en un tubo acelerador donde son acelerados por un campo eléctrico, adquiriendo altas energías. El haz de iones de alta energía pasa a través de scanner verticales y horizontales y es implantando en el substrato semiconductor.

Los iones altamente energéticos que se introducen en el interior substrato del semiconductor pierden su energía mediante colisiones con electrones y núcleos del semiconductor, quedando finalmente en reposo. La distancia total que recorre un ión hasta quedar en reposo se denomina **rango, R** cómo muestra la figura siguiente. La proyección de esta distancia sobre el eje de incidencia se denomina **rango proyectado, R_p**. Puesto que el número de colisiones por unidad de distancia recorrida y la energía perdida en cada colisión son variables aleatorias, por lo que no todos los iones irán a parar a la misma distancia de la superficie, sino que existirá una distribución espacial de los iones que tienen la misma masa y la misma energía inicial. Las fluctuaciones estadísticas en el rango proyectado se denominan **dispersión proyectada, ΔR_p.** Existe también una fluctuación estadística a lo largo del eje perpendicular a la dirección de incidencia, y se denomina **dispersión lateral, $\Delta R_\perp$**.

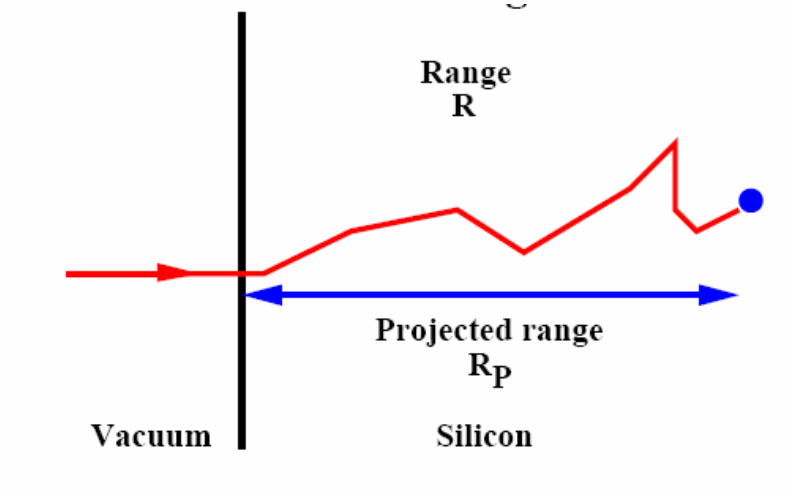


Figura 10.6.2.

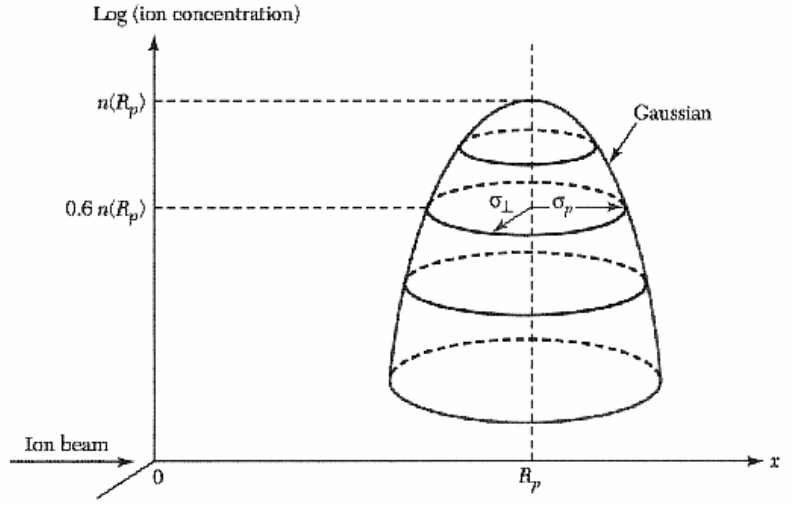


Figura 10.6.3.

A lo largo del eje de incidencia, el perfil de las impurezas implantadas puede aproximarse por una función gaussiana (Ecuación 10.25) tal como muestra la figura 10.6.4.

$$n(x) = \frac{S}{\sqrt{2\pi}\Delta R_p} e^{\frac{-(x-R_p)^2}{2\Delta R_p^2}} \qquad (10.25)$$

donde S es la concentración de iones por unidad de área (dosis).

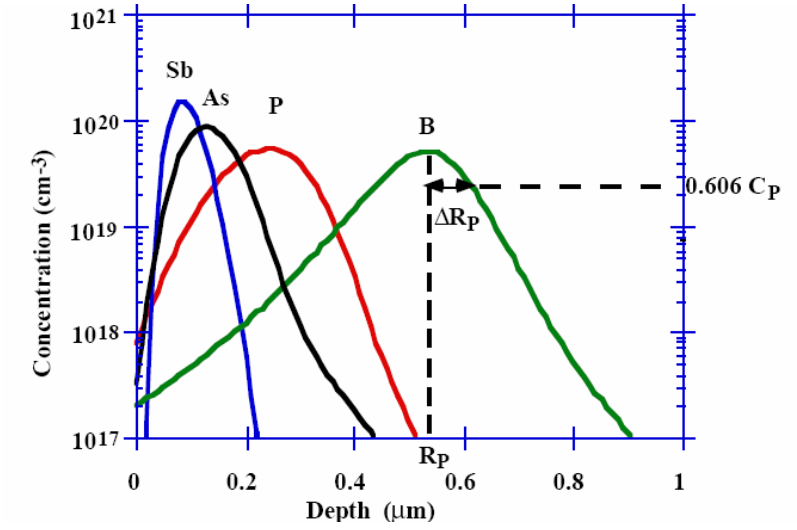


Figura 10.6.4. Perfiles de impurezas implantadas a 200keV.

La concentración en el máximo, C_p, (n(x=R_p), ejemplo 10.1) es igual a:

$$C_p \equiv n(R_p) = \frac{S}{\sqrt{2\pi}\Delta R_p} \qquad (10.26)$$

expresión que nos da una relación muy útil entre la dosis, S, y la concentración de impurezas en el pico, C_p.

La ecuación 10.25 es idéntica a la ecuación 10.16 obtenida en el caso de difusión con cantidad total de dopante constante, excepto que el factor 4Dt es reemplazado ahora por $2\Delta R_p^2$; además, ahora la distribución aparece desplazada a lo largo del eje x una distancia R_p. Por lo tanto, para la difusión el máximo aparece en x=0, mientras que en el caso de implantación iónica el máximo aparece a una distancia R_p de la superficie. La concentración de iones se reduce en un 40% de su valor máximo a una distancia $\pm\Delta R_p$. A lo largo del eje perpendicular al eje de incidencia, la distribución de impurezas sigue también una función gaussiana, por lo que aparecerá también cierta implantación lateral. Sin embargo, ésta es mucho más pequeña que la difusión lateral, cómo veremos próximamente.

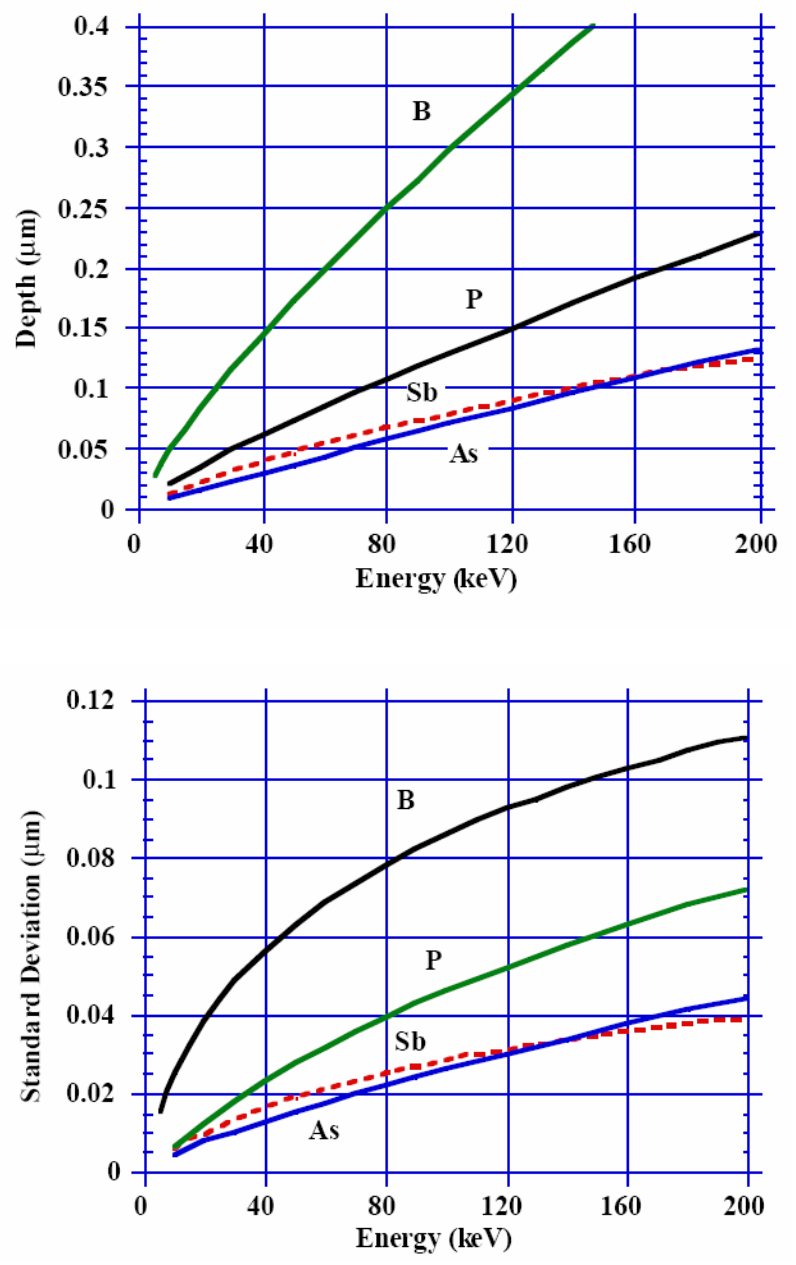


Figura 10.6.5. Rango proyectado, R_p, y desviación típica, ΔR_p, para los dopantes más usuales en silicio, en función de la energía de implantación.

Ejemplo 10.1

Considere un proceso de implantación iónica de boro con una dosis S de $5x10^{14}cm^{-2}$ y una energía de 80keV. Calcule la concentración de boro en el máximo del perfil.

Solución

El perfil de impurezas implantadas para el caso del boro viene descrito en función de la dosis, el rango proyectado y la desviación típica del rango proyectado por la expresión 10.25. El máximo se obtendrá en el punto donde se anule la primera derivada:

$$\frac{dn}{dx} = -\frac{S}{\sqrt{2\pi}\Delta R_p}\frac{\left(x-R_p\right)}{\Delta R_p^2}e^{\frac{-\left(x-R_p\right)^2}{2\Delta R_p^2}} = 0 \qquad (10.27)$$

La derivada (10.27) se anulará para $x=R_p$. Particularizando la ecuación 10.25 para $x=R_p$, se obtiene que el máximo de concentración es igual a:

$$C_p \equiv n(R_p) = \frac{S}{\sqrt{2\pi}\Delta R_p} \qquad (10.28)$$

De acuerdo con la figura 10.6.5, para boro y con una energía inicial de 80keV, $R_p=0.25\mu m$ $\Delta R_p=0.08\mu m$. Por lo tanto el máximo de la distribución es, de acuerdo con 10.28 para una dosis $S=5x10^{14}cm^{-2}$, $C_p=2.49x10^{19}cm^{-3}$.

10.7 Mecanismos de parada

Existen dos mecanismos mediante los cuales los iones pierden su energía a medida que penetran hacia el interior del semiconductor.

El primer mecanismo es mediante la transferencia de energía a los núcleos del substrato (Figura 10.7.1). Esto provoca la desviación del ión incidente de su trayectoria original así como el que el núcleo del semiconductor abandone su posición de equilibrio, provocando dislocaciones en la red.

Si E es la energía de un ión en un punto x de su trayectoria, definimos la potencia de parada nuclear como la energía perdida por el ión por unidad de espacio recorrido debido a colisiones con los núcleos de la red cristalina:

$$S_n(E) = \left.\frac{dE}{dx}\right|_{colisiones} \qquad (10.29)$$

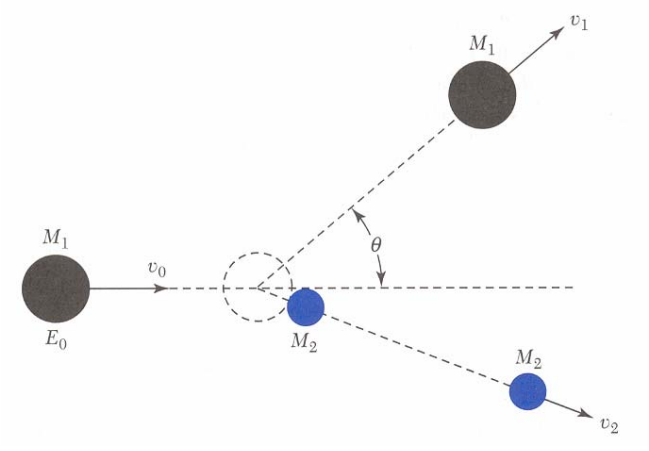


Figura 10.7.1. Diagrama de una colisión entre el ión de impureza (M_1) con energía E_0 y un ión de la red cristalina (M_2) inicialmente en reposo.

El segundo mecanismo mediante los cuales el ión implantado pierde su energía es la interacción culombiana con la nube de electrones que rodean a los átomos de la red cristalina en la que son implantados. Los electrones del átomo de la red cristalina son excitados de esta forma a niveles superiores, pudiendo llegar incluso a la ionización.

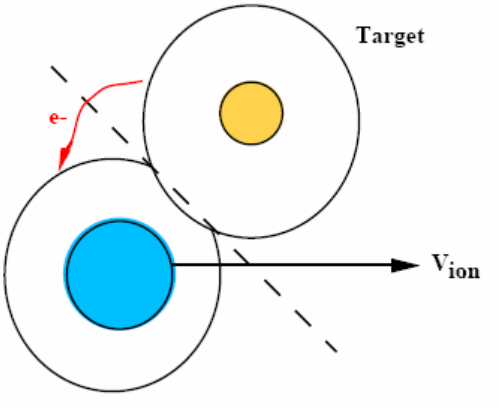


Figura 10.7.2. Diagrama de una colisión electrónica entre el ión implantado y la nube electrónica del substrato implantado.

Al igual que en el caso de las colisiones nucleares, si E es la energía de un ión en un punto x de su trayectoria, definimos la potencia de parada electrónica como la energía perdida por el ión por unidad de espacio recorrido debido a colisiones electrónicas:

$$S_e(E) = \left.\frac{dE}{dx}\right|_e \tag{10.29}$$

La pérdida media de energía con la distancia recorrida puede obtenerse mediante la superposición de los dos mecanismos:

$$\frac{dE}{dx} = \left.\frac{dE}{dx}\right|_n + \left.\frac{dE}{dx}\right|_e = S_n(E) + S_e(E) \tag{10.30}$$

Si un ión posee una energía inicial E_0 la distancia total recorrida hasta perder toda la energía y quedar en reposo, es decir, el rango, R,

$$R = \int_0^R dx = \int_0^{E_0} \frac{dE}{S_n(E) + S_e(E)} \tag{10.31}$$

conocida la energía inicial del ión, E_0, podemos calcular el rango R, supuestas conocidas las funciones $S_n(E)$ y $S_p(E)$.

Estudios detallados del proceso de colisión de iones con átomos del substrato, muestran que $S_n(E)$ aumenta con la energía del ión, para bajos valores de la energía, y alcanza un valor máximo para un valor intermedio de energía. Para altas energías, $S_n(E)$ disminuye con la energía debido a que las partículas a tan alta velocidad no tienen tiempo suficiente de interacción para que se produzca una transferencia efectiva de energía.

En el caso de $S_e(E)$ se ha comprobado que es proporcional a la velocidad de incidencia del ión (o lo que es lo mismo):

$$S_e(E) = k_e\sqrt{E} \tag{10.32}$$

donde k_e es igual a 10^7 $(eV)^{1/2}$/cm para el silicio y $3x10^7$ $(eV)^{1/2}$/cm para el arseniuro de galio.

La figura 10.7.3 muestra los parámetros $S_n(E)$ para arsénico, fósforo y boro en silicio, y $S_s(E)$ para el silicio. A medida que los iones son más pesados la pérdida de energía por colisión con los núcleos es mayor. Se muestra también en la figura los puntos de corte de ambas curvas. Para el boro, el punto de corte es de 10 keV. Esto significa que puesto que el rango normal de energía de implantación es de 30 a 300 keV, en el caso del boro el principal mecanismo de pérdida de energía es el de interacción electrónica. Por el contrario, en el caso de arsénico que tiene un masa atómica elevada, el cruce se produce para 700 keV, por lo que las colisiones con los núcleos son el mecanismo principal de pérdida de energía para los iones implantados.

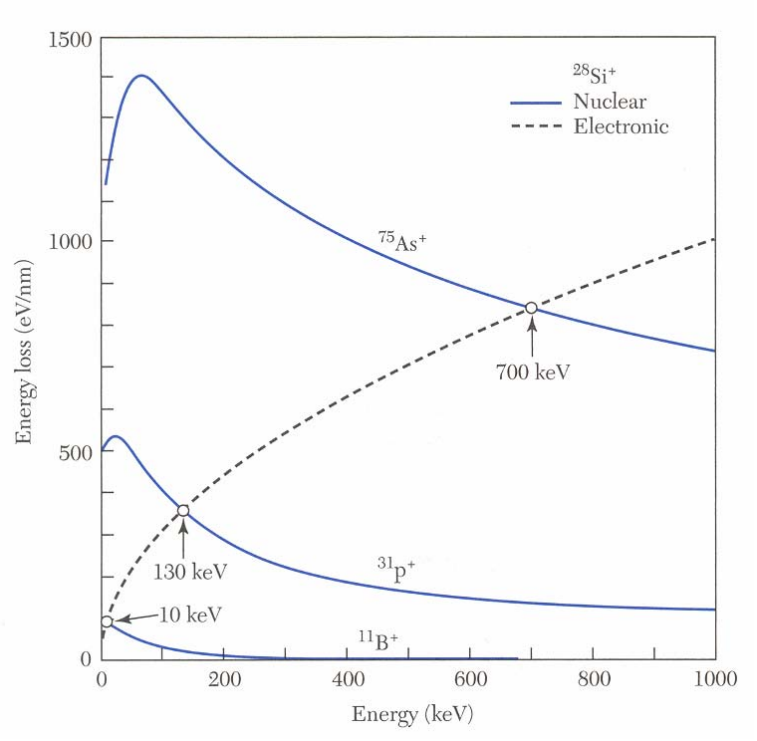


Figura 10.7.3. Potencia de parada nuclear y electrónica para los principales impurezas en silicio, en función de la energía del ión.

Una vez que se conoce $S_n(E)$ y $S_p(E)$, podemos calcular el rango a partir de la ecuación 10.31. Conocido el rango, podemos calcular el rango proyectado y la dispersión proyectada a partir de las expresiones aproximadas siguientes:

$$R_p \approx \frac{R}{1+\frac{M_2}{3M_1}} \tag{10.33}$$

$$\Delta R_p \approx \frac{2}{3}\left(\frac{\sqrt{M_1 M_2}}{M_1+M_2}\right)R_p \tag{10.34}$$

donde M_1 es la masa del ión incidente y M_2 la masa de un átomo de la red cristalina sobre la que se produce la implantación.

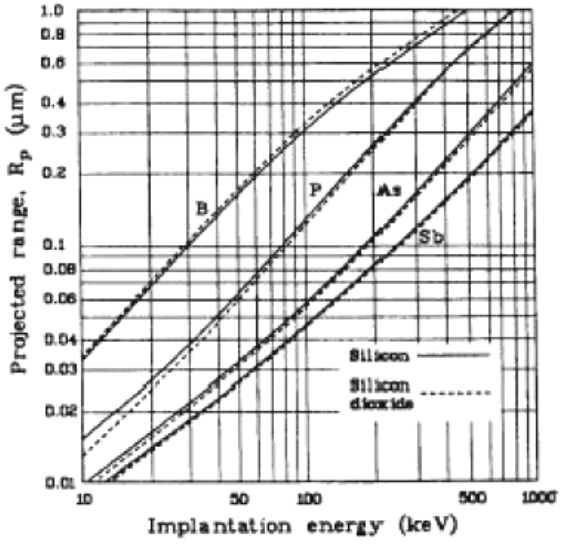


Figura 10.7.4. Rango proyectado en función de la energía de implantación para las principales impurezas implantadas en silicio. Se muestra también, en línea discontinua el rango proyectado en SiO2.

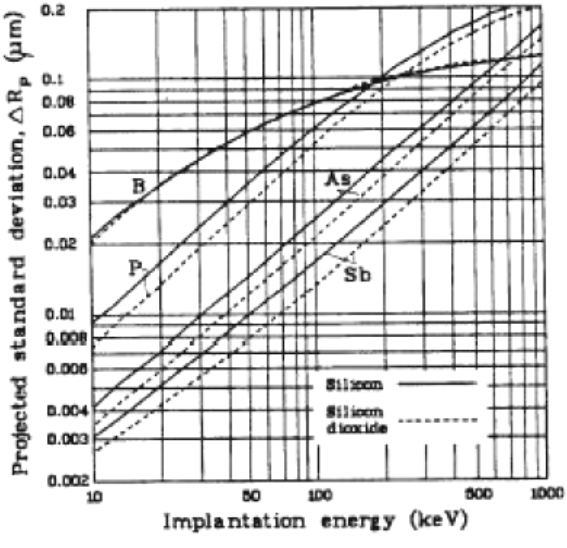


Figura 10.7.5. Desviación del rango proyectado en función de la energía de implantación para las principales impurezas implantadas en silicio. Se muestra también, en línea discontinua el rango proyectado en SiO_2

Por lo que a partir de la energía inicial y la dosis (parámetros tecnológicos que controlan el proceso) hemos encontrado unas expresiones analíticas que nos permiten obtener el perfil de las impurezas implantadas, evaluando la expresión 10.25.

Las figuras 10.7.4 y 10.7.5 muestran los rangos proyectados R_p para arsénico, boro y fósforo en silicio y dióxido de silicio y la desviación típica del rango proyectado, ΔR_p. A medida que mayor es la energía perdida menor es el rango proyectado. En primera aproximación, el rango proyectado varía linealmente con la energía de los iones. También se observa que a igualdad de energía los iones que menor peso atómico recorren una mayor distancia. Se muestra también la dispersión del rango proyectado.

10.8 Reactivación de Impurezas. Desorden y recocido

Cuando los iones energéticos penetran en el substrato semiconductor, pierden su energía tras una serie de colisiones electrónicas y nucleares. La energía perdida en colisiones electrónicas produce excitaciones de los electrones hasta niveles de mayor energía o generación de pares electrón-hueco. Sin embargo las colisiones electrónicas no desplazan a los átomos del semiconductor de su posición de equilibrio en la red. Solamente las colisiones nucleares pueden transferirla suficiente energía a los átomos del semiconductor de forma que el desplazamiento de estos provoque desórdenes en la red. Estos átomos desplazados pueden provocar a su vez una cascada secundaria de desplazamientos en los átomos vecinos dando lugar a lo que se conoce como árbol de desórdenes.

El árbol de desorden provocado por un ión ligero es muy diferente al árbol que producen los iones pesados. La mayoría de la energía perdida por un ión ligero (B^+ en silicio) es debido a colisiones electrónicas que no causan daño en la red. Los electrones pierden su energía a medida que penetran en la red. Únicamente cuando la energía del ión es menor de 10 keV empiezan a ser dominantes las colisiones nucleares, pero entonces el ión se encuentra ya muy lejos de la superficie, por lo que la mayoría de los desórdenes ocurren en la parte final del recorrido.

Por el contrario, para los iones pesados la pérdida de energía se debe principalmente a las colisiones nucleares, de ahí que sea esperable un daño mucho mayor que en el caso de iones ligeros.

Debido al daño causado por la implantación iónica, diferentes parámetros del semiconductor, tales como la movilidad y vida media, sufren una fuerte degradación. Además, la mayoría de los iones implantados no ocupan lugares substitucionales en la red, sino posiciones intersticiales, no siendo por lo tanto eléctricamente activos. Para activar los iones implantados y restaurar los diferentes parámetros del semiconductor dañados por la implantación, el semiconductor se recuece durante un tiempo determinado y a una temperatura apropiada, que depende del tipo de impureza implantada y de la dosis de impurezas implantadas.

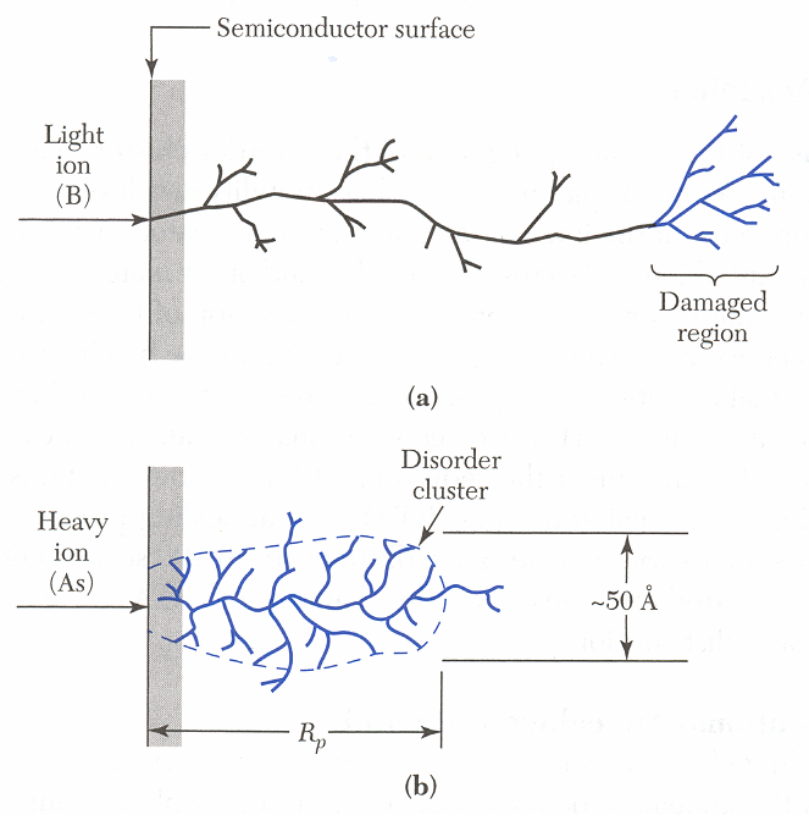


Figura 10.8.1. Árboles de desorden para iones ligeros (boro), y para iones pesados (arsénico).

La figura 10.8.2 muestra la temperatura de recocido para implantaciones de boro y fósforo en silicio. La temperatura de recocido se define, para una dosis de iones dada, como la temperatura a la que el 90% de los iones implantados son activados mediante un recocido de 30 minutos. En el caso del boro, mayores temperaturas de recocido son necesarias para mayores dosis de iones implantados. En el caso del fósforo, el comportamiento es similar al del boro para dosis pequeñas. Sin embargo cuando la dosis es mayor de 10^{15} cm^{-2} el silicio se convierte en amorfo (debido al daño causado por la implantación) y la temperatura de recocido cae por debajo de 600 ºC. Este silicio amorfo puede recristalizarse tomando como semilla la parte de silicio cristalino que queda por debajo y que no ha sido dañada por la implantación. A este fenómeno se le conoce como **epitaxia en fase-sólida**, con la que se consiguen altas velocidades de crecimiento (100 Å/min) a relativamente bajas temperaturas.

Durante el recocido, el perfil implantado puede ser ensanchado por difusión. El perfil inicial de los iones implantados es una distribución gaussiana. El perfil obtenido mediante una difusión con cantidad constante de dopante es también gaussiano.

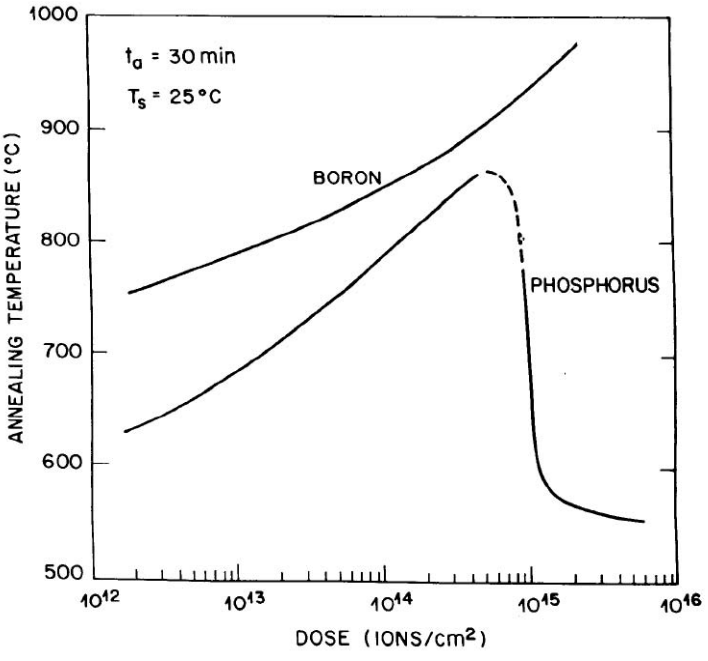


Figura 10.8.2. Temperatura de recocido para boro y fósforo en silicio en función de la dosis implantada. Tiempo de recocido=30 min.

Por lo tanto, el perfil después del recocido viene dado por:

$$n(x) = \frac{S}{\sqrt{2\pi}\Delta R_p} e^{-\frac{(x-R_p)^2}{2(\Delta R_p^2 + 2Dt)}} \qquad (10.35)$$

donde t es el tiempo de recocido y D el coeficiente de difusión de las impurezas a la temperatura de recocido. Como se observa, en un recocido en horno convencional se produce un ensanchamiento importante debido al tiempo elevado de recocido. Para evitar esto, lo que se hace es implantar iones inertes que conviertan a la superficie del silicio en amorfa, y proceder después a una epitaxia en fase sólida. Recientemente se han utilizado otras fuentes energéticas para llevar a cabo el recocido, como por ejemplo recocido con haces de electrones. Los resultados se muestran en la figura 10.8.4.

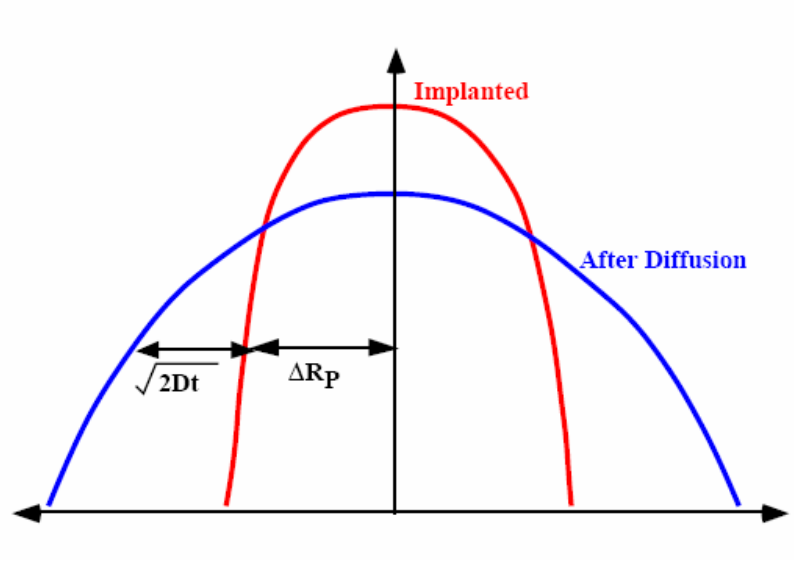


Figura 10.8.3. Redistribución del perfil de impurezas implantadas tras el recocido térmico para reactivación de las impurezas.

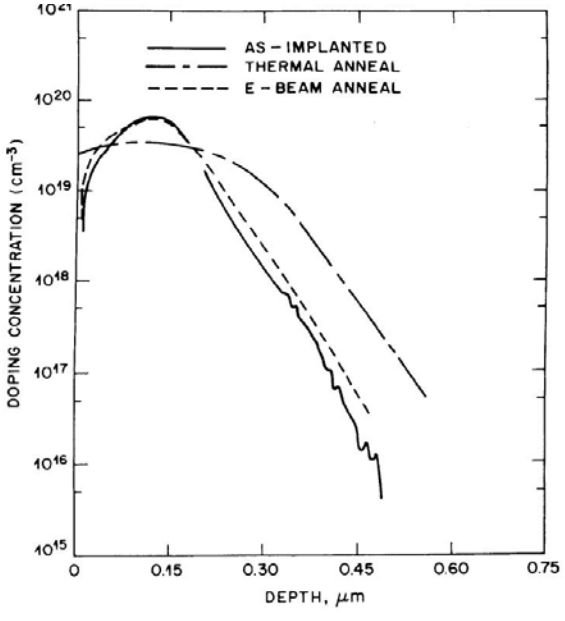


Figura 10.8.4. Perfil de impurezas implantadas (línea continua) y obtenido tras un proceso de reactivación de impurezas por recocido térmico o un proceso de reactivación de impurezas por bombardeo con electrones energéticos.

Máscara para la implantación

Las figuras 10.7.4 y 10.7.5 muestran el rango proyectado de las principales impurezas en silicio y dióxido de silicio. Se observa que el rango proyectado en silicio y en dióxido de silicio prácticamente coincide, lo que supone una diferencia apreciable con el caso de la difusión en donde vimos que el dióxido de silicio se comporta como una máscara efectiva para la difusión de impurezas, es decir, por término medio, la longitud de difusión en el silicio es un orden de magnitud mayor que la longitud de difusión en el SiO_2 en igualdad de condiciones. No ocurre lo mismo en el caso de la implantación, como acabamos de mencionar. Para que la máscara sea efectiva, su espesor, x_m, debe ser tal que:

$$C^*(x_m) = C_p^* e^{-\frac{\left(x_m - R_p^*\right)^2}{2\left(\Delta R_p^*\right)^2}} \leq C_{substrato} \tag{10.36}$$

siendo R_p^* y ΔR_p^* el rango proyectado y la desviación típica del rango proyectado en la máscara y $C_{substrato}$ la concentración inicial (residual) de impurezas en el substrato.

Despejando de la ecuación 10.36, se obtiene que el espesor de la máscara debe ser mayor de:

$$x_m = R_p^* + \Delta R_p^* \sqrt{2 \ln\left(\frac{C_p^*}{C_B}\right)} = R_p^* + m\Delta R_p^* \tag{10.37}$$

La dosis de impurezas implantadas que atraviesan la máscara y alcanzan el substrato viene dada por:

$$S_p = \frac{S}{\sqrt{2\pi}\Delta R_p^*} \int_{x_m}^{\infty} e^{-\frac{\left(x_m - R_p^*\right)^2}{2\left(\Delta R_p^*\right)^2}} dx = \frac{S}{2} erfc\left(\frac{x_m - R_p^*}{\sqrt{2}\Delta R_p^*}\right) \tag{10.38}$$

siendo S la dosis total de las impurezas implantadas.

Implantación lateral

La figura 10.9.1 muestra el perfil bidimensional de una implantación de iónica de fósforo en silicio:

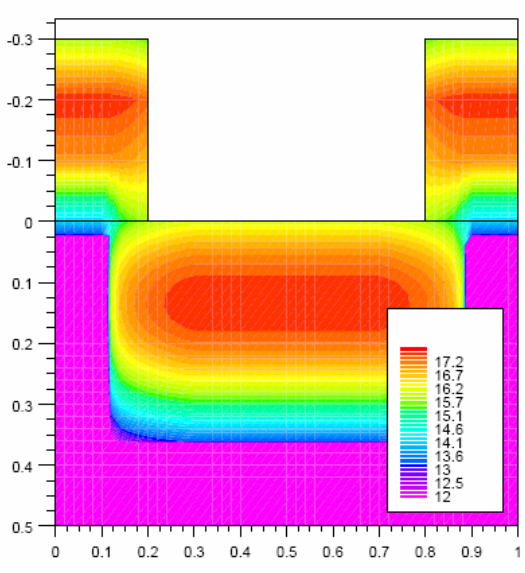


Figura 10.9.1. Perfil bidimensional de impurezas implantadas.

En el extremo de la máscara se observa que al igual que la difusión también se produce una pequeña penetración de impurezas por debajo de la máscara. Si la máscara es suficientemente gruesa, de manera que las impurezas no la atraviesan (como es el caso) el perfil de impurezas en el extremo de la máscara viene determinado por la desviación lateral, $\Delta R_{\perp}$, que es mucho menor que la difusión lateral.

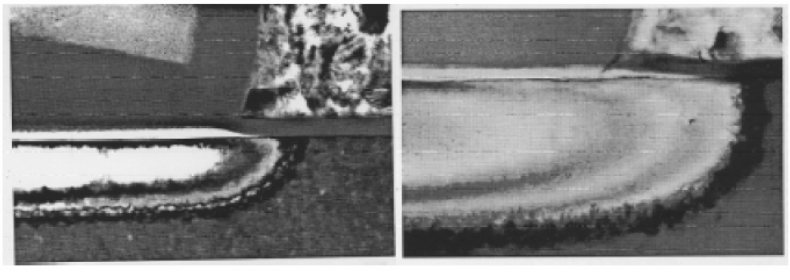


Figura 10.9.2. Fotografía que muestra el perfil resultante de la implantación de arsénico a 35keV y 120keV en el extremo de la máscara de polysilicio.

Acanalamiento

El substrato sobre el que se realiza la implantación suele ser una red cristalina, es decir, un conjunto de átomos ordenados espacialmente siguiendo un determinado patrón. Existen determinadas direcciones cristalográficas privilegiadas en la que existen grandes huecos, grandes intersticios, como muestra la figura 10.9.3.

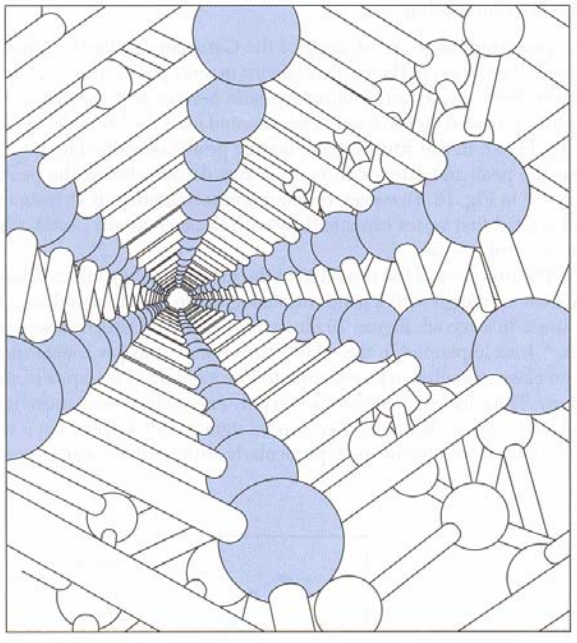


Figura 10.9.3. Sucesión de intersticios en una red cristalina.

Si los iones implantados lo hacen siguiendo una de estas direcciones pueden penetrar en el interior del cristal sin apenas sufrir pérdida de energía. En realidad, el único mecanismo de pérdida de energía en este caso son las colisiones electrónicas. En consecuencia, el rango proyectado es mucho mayor de lo esperado.

Para evitar este comportamiento, se pueden adoptar alguna de las soluciones mostradas en la figura 10.9.4:

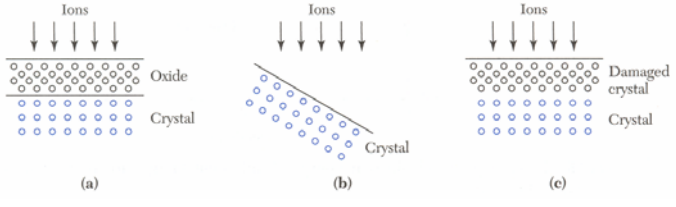


Figura 10.9.4: Alternativas para evitar el acanalamiento.

RESUMEN

Para controlar la conductividad localmente en una muestra semiconductora es necesario impurificar el semiconductor con impurezas en posición susbstitucional. Se han estudiado las diferentes técnicas de contaminación controlada (difusión e implantación iónica) resaltando las ventajas e inconvenientes de cada proceso. Se han obtenido expresiones analíticas para ambos procesos, que proporcionan el perfil de impurezas en función de los parámetros tecnológicos de cada proceso.

CUESTIONES Y PROBLEMAS

1. Se realiza una implantación de fósforo sobre silicio con una dosis $D=2x10^{15}cm^{-2}$ a una energía de 40 KeV. Obtenga el perfil de impurezas. A continuación se realiza un recocido térmico a 900 ºC. Calcule el tiempo que debe transcurrir para que una profundidad de 0.2µm se alcance una concentración de $10^{17}cm^{-3}$.
2. Calcule el espesor de oxido necesario para bloquear el 98% de una implantación de As a 40 KeV.

REFERENCIAS

[1] S. Sze. *VLSI Technology*, Ed. Mcgraw-Hill.

[2] J.D.Plummer, M.D.Deal, P.B.Griffin, *Silicon VLSI Technology*, Ed.Prentice Hall.

[3] Chang and S. Sze, *ULSI Technology*, Ed. Mcgraw-Hill.

[4] Streetman and Banerjee, *Solid State Electronic Devices*, Prentice Hall,Fifth edition, 2000.

[5] Glosario: http://semiconductorglossary.com/

11

Capítulo

Crecimiento y Deposición de Películas Delgadas

Horno de Oxidación

Índice

Objetivos

- Describir el papel de diferentes películas delgadas de distintos materiales en el funcionamiento y fabricación de los dispositivos electrónicos.
- Describir las técnicas de fabricación y propiedades de los óxidos térmicos o nativos.
- Estudio de las técnicas de deposición de dieléctricos, especialmente de óxidos no-nativos.
- Importancia del silicio policristalino en la fabricación de dispositivos electrónicos. Técnicas de deposición.
- Describir las técnicas de deposición de películas metálicas.

Palabras Clave

Óxidos nativos
Oxidación
Oxidación seca
Oxidación húmeda
Cinética de oxidación
Óxidos delgados
Silano
CVD
LPCVD
PCVD
TEOS
Nitruro de silicio
LOCOS
Polisilico
PVD
Metalización
Sputtering
Metales refractarios
Siliciuros
Electromigración

11.1 Introducción

Para la fabricación tanto de dispositivos discretos cómo de circuitos integrados es necesario construir diferentes láminas delgadas. Estas láminas delgadas pueden ser clasificadas en cuatro tipos: óxidos térmicos (óxidos crecidos térmicamente), capas de aislantes, silicio policristalino y películas metálicas. La figura 11.1.1 muestra un gráfico esquemático de un MOSFET canal-n en el que se observan los cuatro tipos de láminas delgadas anteriores

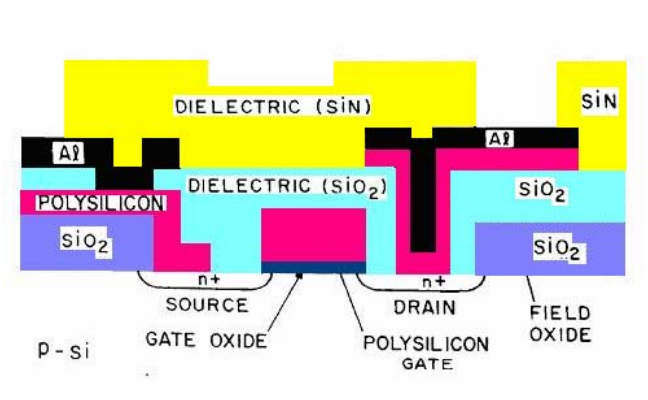


Figura 11.1.1. Esquema de un MOSFET canal N.

Un primer representante de las películas delgadas del grupo de óxidos crecidos térmicamente es el **óxido de puerta (gate oxide)** bajo el que puede formarse un canal conductor entre la fuente y el drenador. Una capa relacionada con la anterior es el **óxido de campo (field oxide)** utilizado para aislar los diferentes dispositivos de un mismo circuito integrado. Ambos óxidos (óxido de puerta y óxido de campo) son crecidos mediante un proceso de oxidación térmica, ya que es el único proceso que proporciona óxidos de buena calidad.

Las capas de dieléctricos tales como la **capa depositada de dióxido de silicio** y **nitruro de silicio** se utilizan como aislantes entre las diferentes capas conductoras, o bien como máscaras para la difusión o implantación iónica, para prevenir la pérdida de dopantes, y para pasivación y prevención de los dispositivos de impurezas perjudiciales. El **silicio policristalino** o **polisilicio** se utiliza como electrodo de puerta en dispositivos MOS, como material conductor para dispositivos con varios niveles de metalización, y como material de contacto para dispositivos con uniones poco profundas. Las láminas metálicas de aluminio y siliciuros se emplean para formar interconexiones de baja resistencia, contactos óhmicos, y contactos rectificadores metal-semiconductor.

11.2 Oxidación térmica

Los semiconductores pueden ser oxidados mediante diferentes técnicas (oxidación térmica, anodización electroquímica, reacción con plasmas). De todas ellas, la más importante con diferencia, para dispositivos de silicio es la oxidación térmica. Para el arseniuro de galio la oxidación térmica da lugar a capas no estequiométricas que contienen óxidos de galio y arsénico e iones de arsénico. Estos óxidos proporcionan un aislamiento eléctrico pobre. Por consiguiente estos óxidos se utilizan raramente para dispositivos de arseniuro de galio.

La figura 11.2.1 muestra el montaje básico para un proceso de oxidación térmica. Éste consiste en un horno calentado por resistencia, un tubo cilíndrico de cuarzo fundido que contiene a las obleas de silicio verticalmente y una fuente bien de oxígeno puro seco, bien de vapor de agua. En el extremo de carga del horno se mantiene un flujo de aire filtrado que elimina el polvo y minimiza la contaminación de las obleas. La temperatura de oxidación está normalmente en el rango de 900 a 1200ºC.

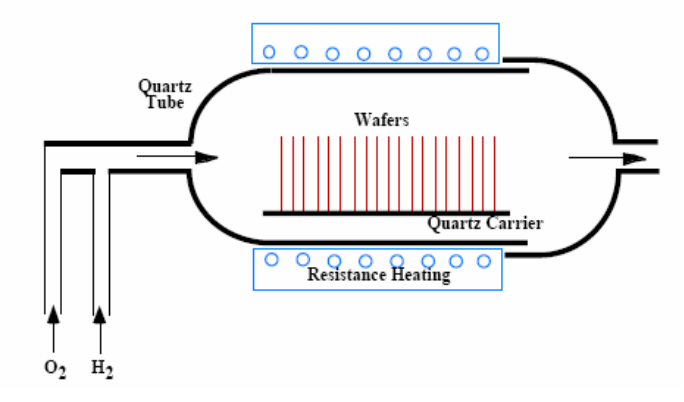


Figura 11.2.1. Montaje básico para el proceso de oxidación.

Para que pueda producirse el proceso de oxidación es necesario poner el silicio en contacto con las especies oxidantes, oxígeno o vapor de agua. A medida que se forma el óxido, las especies oxidantes tienen que atravesar el óxido previamente crecido para alcanzar el silicio de la interfase, tal como muestra la figura 11.2.2. Una vez en contacto, oxidantes y silicio, se produce la oxidación cuyo resultado es la formación de dióxido de silicio, SiO_2, y la desaparición del silicio de la oblea semiconductora. Por lo tanto la oxidación térmica, involucra a su vez dos procesos

i) difusión de oxidantes a través del óxido

ii) reacción química de oxidación.

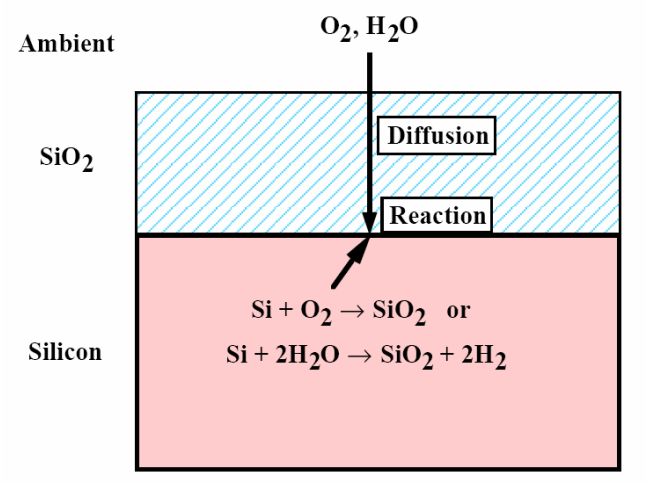


Figura 11.2.2. El proceso de oxidación térmica necesita que los oxidantes atraviesen el óxido previamente crecido para alcanzar el silicio del substrato y se pueda producir la reacción química.

Durante el proceso de oxidación la interface silicio-dióxido de silicio se desplaza hacia el interior del silicio, de manera que para crecer una capa de óxido de espesor x, se consume una capa de silicio de espesor 0.44x, (Figura 11.2.3) lo que produce cierta tensión en la estructura que puede provocar dislocaciones (Figura 11.2.4).

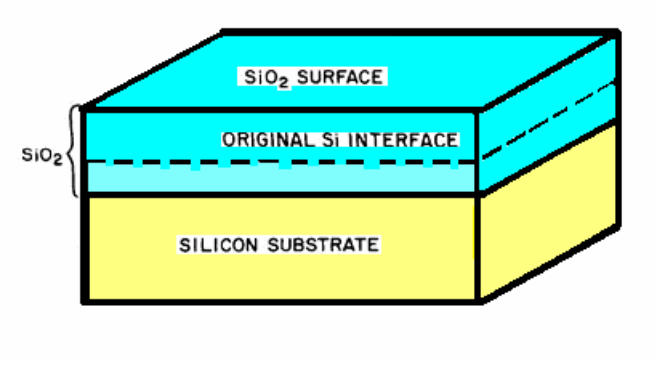


Figura 11.2.3. Al formarse SiO_2 desaparece Si del substrato. El mayor volumen del SiO_2 formado hace que para crecer una capa de óxido de espesor *x*, se consume una capa de silicio de espesor 0.44*x*.

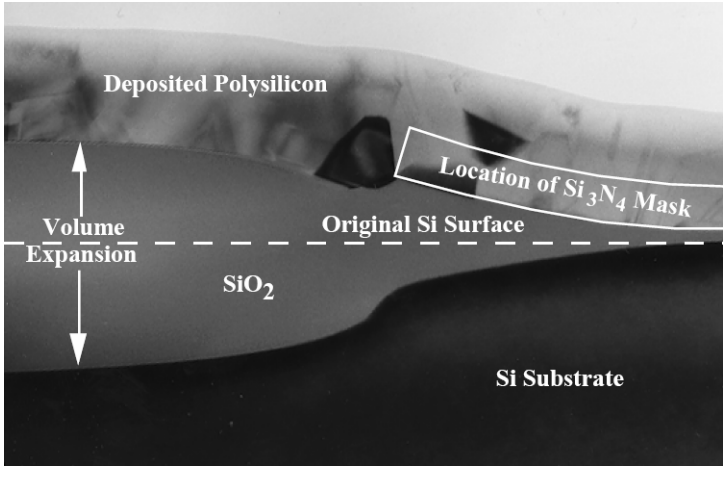


Figura 11.2.4. El mayor volumen del SiO_2 provoca la aparición de tensiones en la estructura de silicio, y superficies no planas.

Las siguientes reacciones químicas describen el proceso de oxidación térmica del silicio con oxígeno o con vapor de agua:

$$Si\,(s) + O_2\,(g) \rightarrow SiO_2\,(s) \qquad (11.1)$$

$$Si\,(s) + 2H_2O\,(g) \rightarrow SiO_2\,(s) + 2H_2 \qquad (11.2)$$

Cuando se utiliza como oxidante oxígeno molecular, se dice que la oxidación se realiza en ambiente seco (ecuación 11.1) mientras que si se emplea vapor de agua, se dice que la oxidación se realiza en ambiente húmedo (ecuación 11.2).

La unidad básica estructural del dióxido de silicio térmicamente crecido es un ión de silicio rodeado por cuatro átomos de oxígeno formando un tetraedro (Fig11.2.5 a). Estos tetraedros se unen unos a otros por los extremos compartiendo un átomo de oxígeno de distintas formas que dan lugar a diferentes estructuras de dióxido de silicio (sílice). La sílice tiene varias formas cristalinas (por ejemplo el cuarzo) y una forma amorfa. Cuando el dióxido de silicio se crece térmicamente se obtiene una estructura amorfa. La estructura amorfa es menos densa que la cristalina debido a la mayor cantidad de huecos existentes. Además estos huecos permiten que los átomos de impurezas (por ejemplo sodio) se difundan de forma relativamente fácil a través de la capa de dióxido de silicio.

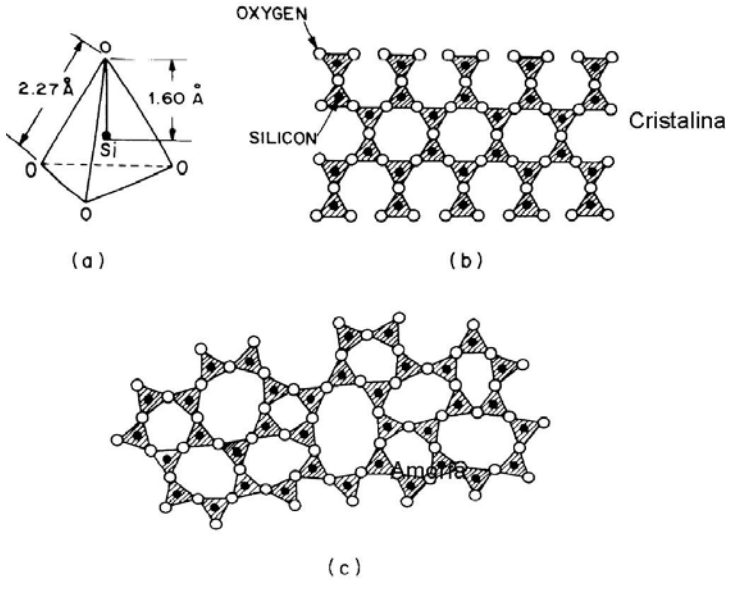


Figura 11.2.5. (a) Unidad estructural del SiO_2. (b) Sílice. Estructura cristalina. (c) Sílice. Estructura amorfa.

Cinética de oxidación (Modelo de Deal-Grove)

Necesitamos desarrollar un modelo que nos proporcione el espesor de la capa de óxido en función de las variables tecnológicas del proceso, esto es, el tiempo y la temperatura. Para ello, teniendo en cuenta que la formación de SiO_2 involucra la participación de dos procesos, difusión de oxidantes a través del óxido crecido, y reacción de oxidación, consideremos la figura siguiente11.2.6.

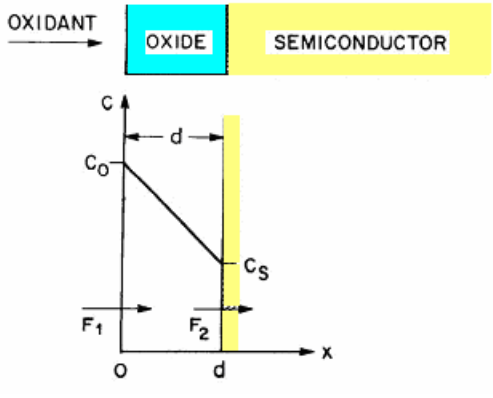


Figura 11.2.6. Esquema para la obtención del modelo de oxidación de Deal-Grove.

La superficie del óxido se encuentra en contacto con los oxidantes (oxígeno o vapor de agua) en una concentración C_0, que a una

temperatura de 1000 ºC y a una presión de 1 atm, resulta ser de C_0= $5.2x10^{16}$ moléculas /cm^3 para el oxígeno seco y de C_0= $3x10^{19}$ moléculas /cm^3 para el vapor de agua. Las especies oxidantes se difunden a través del óxido ya crecido, de manera que a la superficie del silicio (interface silicio/dióxido de silicio) llega una concentración C_s. El flujo de moléculas que atraviesa el óxido, F_1, puede ponerse cómo:

$$F_1 = D\frac{dC}{dx} \approx D\frac{C_0 - C_s}{x} \tag{11.3}$$

donde D es el coeficiente de difusión y x el espesor del óxido ya crecido. En la superficie del silicio los oxidantes reaccionan con este, siguiendo alguna de las reacciones químicas detalladas anteriormente. Suponiendo que la velocidad de reacción es proporcional a la concentración de especies oxidantes en la superficie del silicio, el flujo F_2, de moléculas de oxidante que desaparecen en la interface entre Si y SiO_2, viene dado por:

$$F_2 = kC_s \tag{11.4}$$

donde k es la constante de velocidad de la oxidación. En el estado estacionario, ambos flujos deben ser iguales, $F_1 = F_2 \equiv F$, de ahí que finalmente se obtenga:

$$F = \frac{DC_0}{x + \left(\frac{D}{k}\right)} \tag{11.5}$$

reacción de las especies oxidantes con silicio forma dióxido de silicio. Sea C_1 el número de moléculas de especies oxidantes en una unidad de volumen del óxido. Por cada cm^3 de óxido hay $2.2x10^{22}$ moléculas de dióxido de silicio. Por cada molécula de SiO_2 se necesita una molécula de oxígeno (O_2) y dos moléculas de (H_2O). Por lo tanto, C_1=$2.2x10^{22}$ cm^{-3} para la oxidación en oxígeno seco, mientras que C_1=$4.4x10^{22}$ cm^{-3} en oxidación con vapor de agua. La velocidad de crecimiento de la capa de óxido viene dada por el cociente entre la velocidad de moléculas que llegan a la interface Si-SiO_2 y las que se incorporan a la capa de óxido:

$$\frac{dx}{dt} = \frac{F}{C_1} = \frac{D\frac{C_0}{C_1}}{x + \left(\frac{D}{k}\right)} \tag{11.6}$$

Esta ecuación diferencial puede resolverse sujeta a las condiciones iniciales, $x(0)=d_0$, donde d_0 es el espesor inicial del óxido. La solución a la ecuación diferencial anterior viene dada por:

$$x^2 + 2\frac{D}{k}x = \frac{2DC_0}{C_1}(t+\tau) \tag{11.7}$$

donde por definición:

$$\tau \equiv \frac{\left(d_0^2 + \frac{2Dd_0}{k}\right)C_1}{2DC_0} \tag{11.8}$$

Resolviendo la ecuación de segundo grado (11.7) podemos calcular el espesor del óxido tras un tiempo de oxidación t:

$$x = \frac{D}{k}\left(\sqrt{1 + \frac{2C_0k^2(t+\tau)}{DC_1}} - 1\right) \tag{11.9}$$

La expresión 11.9 nos proporciona el valor del espesor de óxido en función del tiempo, una vez conocidos D y k, que son función de la temperatura y del tipo de oxidación (seca o húmeda) utilizada. Por lo tanto, proporciona el modelo que nos planteamos en un principio. Para obtener una mayor información, consideremos los límites asintóticos de esta expresión. Para valores muy pequeños del tiempo $t+\tau$, podemos usar la aproximación $\sqrt{1+x} \approx 1 + \frac{x}{2}$, con lo que el espesor de óxido queda:

$$x = \frac{kC_0}{C_1}(t+\tau) \tag{11.10}$$

Es decir, el espesor de óxido depende linealmente del tiempo, y viene limitado por la velocidad de oxidación (no depende del coeficiente de difusión de las especies oxidantes a través del óxido previamente crecido).

Para valores grandes del tiempo $t+\tau$, el espesor de óxido puede ponerse como

$$x = \left(\sqrt{\frac{2C_0D(t+\tau)}{C_1}}\right) \tag{11.11}$$

Es decir, el espesor de óxido aumenta con la raíz cuadrada del tiempo, y además la velocidad de crecimiento está limitada por la difusión de los oxidantes a través del óxido crecido (no hay dependencia con la velocidad de oxidación).

De acuerdo con las expresiones anteriores, en las primeras etapas de crecimiento del óxido, cuando la reacción en la interface Si-SiO_2 es el factor que limita la velocidad de crecimiento, el espesor del óxido varía

linealmente con el tiempo. Cuando el espesor de óxido comienza a ser mayor, las especies oxidantes deben difundirse a través del óxido ya crecido para alcanzar la interface, de manera, que la velocidad de crecimiento viene limitada por la difusión de los oxidantes a través del óxido. En este caso, el espesor de óxido es proporcional a la raíz cuadrada del tiempo de crecimiento, lo que implica una velocidad de crecimiento parabólica.

Con las definiciones siguientes,

$$A \equiv \frac{2D}{k} \tag{11.12}$$

$$B \equiv \frac{2DC_0}{C_1} \tag{11.13}$$

Las expresiones 11.9 a 11.11quedan ahora:

$$x^2 + Ax = B(t + \tau) \tag{11.14}$$

Región lineal:

$$x = \frac{B}{A}(t + \tau) \tag{11.15}$$

Región parabólica:

$$x^2 = B(t + \tau) \tag{11.16}$$

Por esta razón, el término B/A se denomina constante de velocidad lineal, mientras que el término B se denomina constante de velocidad parabólica. Las figuras 11.2.7 y 11.2.8 muestran los valores experimentales de estos coeficientes para el silicio en función de la temperatura:

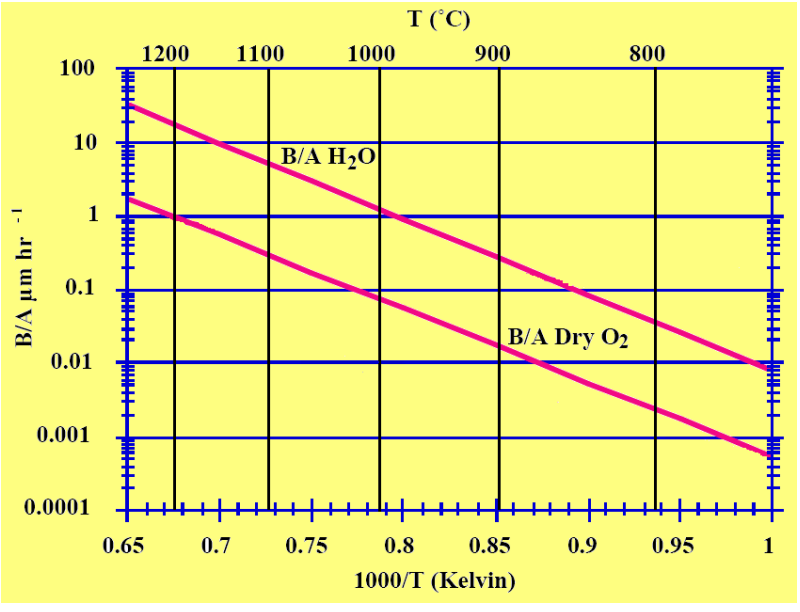


Figura 11.2.7. Constante de velocidad lineal en función del inverso de la temperatura.

En las figuras se aprecia que la constante de velocidad lineal depende de la orientación del cristal. Este comportamiento se debe a que este coeficiente está relacionado con la velocidad de incorporación de los átomos del reactivo a la red, lo que a su vez depende de la densidad de átomos de silicio en la superficie. Como la densidad de átomos de silicio es mayor en la superficie (111) que en la (100), la constante de crecimiento lineal es mayor para una superficie 111 que para una superficie 100. Por otro lado la constante parabólica no depende de la orientación de la superficie. Este resultado es esperable puesto que en este caso la velocidad de crecimiento está controlada por la difusión de los reactivos a través del óxido ya crecido.

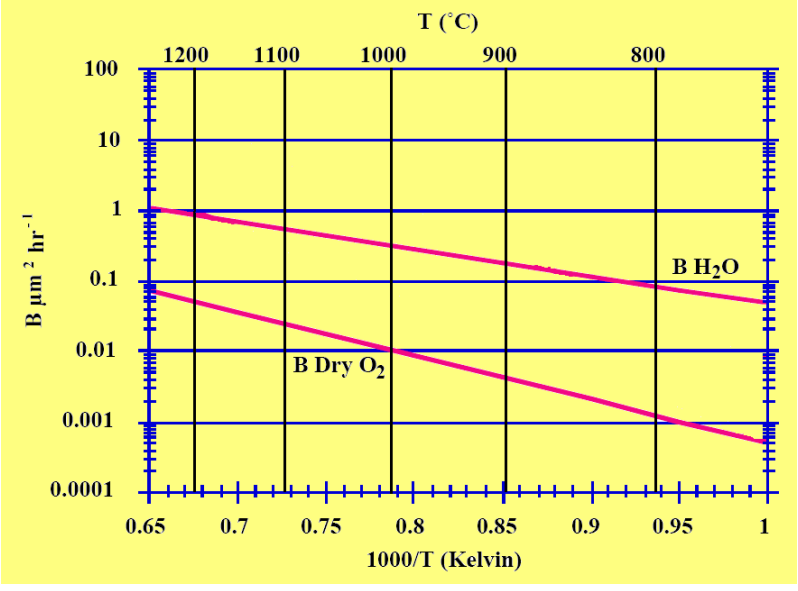


Figura 11.2.8. Constante de velocidad parabólica para oxidación seca y oxidación húmeda en función del inverso de la temperatura.

Los óxidos crecidos en ambiente seco tienen mejores propiedades eléctricas que los crecidos en ambiente de vapor de agua. Sin embargo, se necesitan tiempos considerablemente mayores para obtener el mismo espesor de óxido. Para óxidos delgados, como el óxido de puerta en un transistor MOSFET, se utiliza oxidación seca. Sin embargo para óxidos más gruesos, como los óxidos de campo, se utiliza oxidación en ambiente húmedo.

El modelo descrito anteriormente reproduce adecuadamente los resultados experimentales para la oxidación en ambiente húmedo, esto es, el espesor de óxido previstos por la ecuación 11-14 para un tiempo de oxidación determinado coincide muy aproximadamente con el espesor que se obtiene experimentalmente si se trata de una oxidación en ambiente húmedo, esto es, usando H_2O como oxidante. Sin embargo, para el caso de la oxidación en ambiente seco, usando O_2 como oxidante, el modelo que hemos desarrollado únicamente reproduce el comportamiento experimental para espesores de óxido crecido por encima de 200 Å, colocando como condición inicial d_0=200Å (espesor de óxido inicial) aun cuando no haya ningún óxido inicialmente, ó el espesor de éste sea menor de 200Å.. Experimentalmente se observa que en el caso de la oxidación seca se produce un crecimiento muy rápido hasta los primeros 200Å. Se han propuesto varias teorías para tratar de explicar este fenómeno, pero ninguna de ellas ha sido plenamente aceptada.

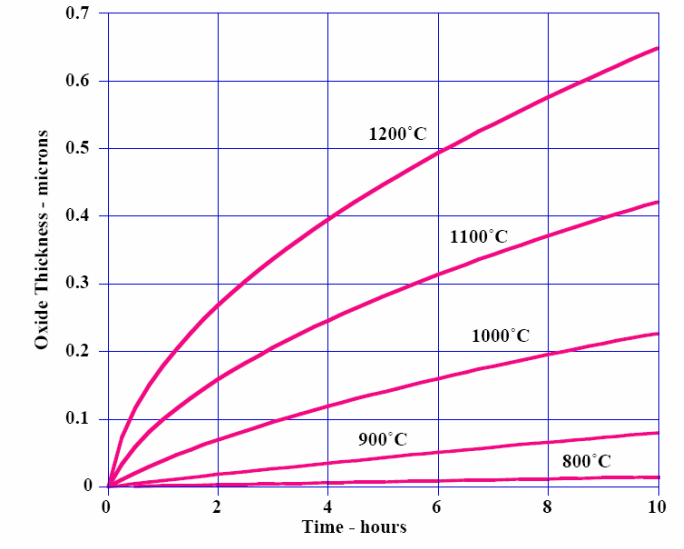


Figura 11.2.9. Espesor de óxido crecido en ambiente seco en función del tiempo de oxidación a diferentes temperaturas.

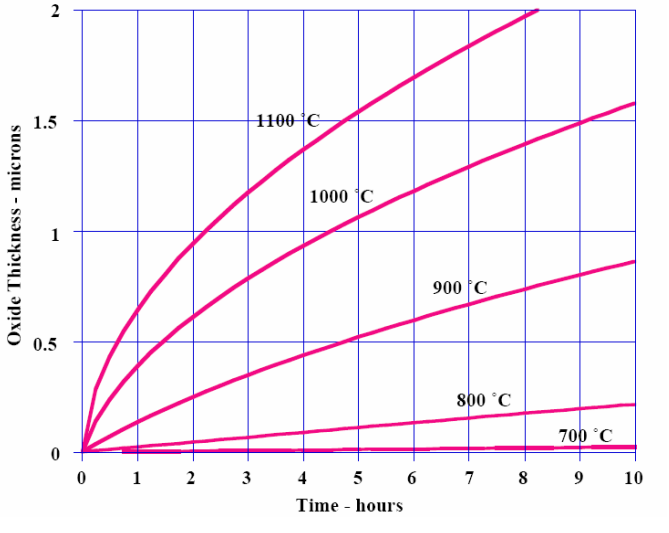


Figura 11.2.10. Espesor de óxido crecido en ambiente húmedo en función del tiempo de oxidación a diferentes temperaturas.

Ejemplo

Se oxida una muestra de silicio en ambiente seco (O_2) durante una hora a 1200ºC. (a) ¿Cual es el espesor de óxido crecido?

(b) ¿Cuánto tiempo adicional es necesario para crecer 0.1µm más de espesor en ambiente húmedo a 1200ºC?

Solución

(a) De la figura 11.2.7 a temperatura de 1200ºC y para una oxidación seca se obtiene que la constante de velocidad lineal B/A es igual a 1 µm/hr mientras que la constante de velocidad parabólica B es igual a 0.052 $\mu m^2/hr$, por lo tanto, la constante A=0.052µm. Para calcular el espesor de óxido necesitamos calcular τ, que es función del espesor de óxido inicial que en el caso de oxidación seca es d_0=200 Å.

$$\tau = \frac{d_0(d_0 + A)}{B} = 0.028\ hr \qquad (11.17)$$

Con estos parámetros y usando la ecuación

$$x^2 + Ax = B(t + \tau) \qquad (11.18)$$

se obtiene un espesor de óxido de **x=0.206µm.**

(b) De la misma forma que en el caso anterior, lo primero es obtener de las gráficas los valores de las constantes de oxidación lineal y parabólica, pero ahora para una oxidación húmeda: B/A=16.7 µm/hr, B=0.839$\mu m^2/hr$ por lo que A=0.05µm. Puesto que se trata de una oxidación húmeda usamos como condición inicial el espesor inicial de óxido, que es el resultado del apartado anterior, esto es, d_0=0.206µm, con lo que τ=0.063hr. Queremos añadir un espesor de óxido de 0.1µm a lo que ya teníamos, esto es, d_0=0.206µm, por lo que el espesor final es x=0.306 µm. Con estos parámetros en la ecuación 11.18 se obtiene que el tiempo necesario es t=0.067hr=4 min

Crecimiento de óxidos delgados

Anteriormente hemos visto que para oxidación seca hay una primera etapa de oxidación rápida que proporciona un óxido inicial de espesor d_0 (~200 Å). Por lo tanto el modelo simple que acabamos de estudiar no es válido para estudiar procesos de oxidación en ambiente seco cuando el espesor del óxido es menor de 200 Å. Por otro lado, con las reglas de escalado, la capacidad de producir óxidos delgados ha crecido espectacularmente en los últimos años.

Desde un punto de vista práctico, el crecimiento de óxidos delgados debe ser lo suficientemente lento para garantizar uniformidad y reproductibilidad. La figura 11.2.11 muestra el espesor del óxido en función del tiempo de oxidación para diferentes temperaturas y presión parcial de oxígeno.

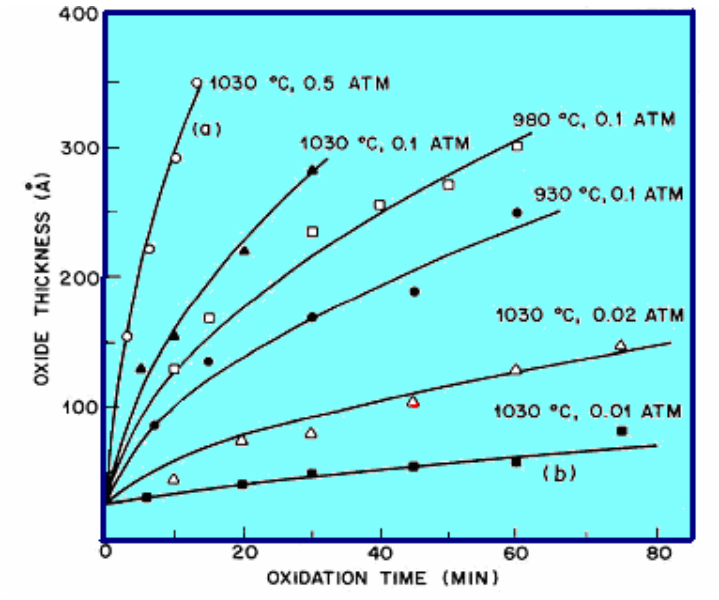


Figura 11.2.11. Espesores de óxido en función del tiempo para diferentes combinaciones de

Reduciendo la presión parcial de oxígeno y la temperatura es posible crecer óxidos delgados suficientemente despacio para garantizar la reproductibilidad. En la figura 11.2.11 se observa también que aumentando la presión parcial de oxígeno se consigue un aumento importante de la velocidad de crecimiento. El crecimiento de óxidos gruesos a altas presiones permite reducir la temperatura del proceso, lo que minimiza el movimiento de impurezas previamente implantadas y la difusión lateral.

Efectos de las impurezas sobre la oxidación

Tanto los valores de las constantes de velocidad parabólica como lineal son sensibles a las impurezas existentes tanto en el gas oxidante como en el substrato de silicio.

-Agua: La existencia de moléculas de agua en el gas oxidante en un proceso de oxidación seca produce un aumento importante en la velocidad de crecimiento.

-Cloro: En los modernos circuitos integrados, el cloro se introduce en el ambiente de oxidación para mejorar la calidad del óxido y las propiedades de la interface Si-SiO_2. La acción del cloro tiene lugar al eliminar ciertas impurezas de la superficie del silicio mediante la formación de cloruros volátiles.

-Dopantes básicos: Los dopantes básicos del silicio, (boro y fósforo) pueden mejorar el comportamiento de la oxidación cuando se encuentran a concentraciones elevadas. Las impurezas de los dopantes son redistribuidas a medida que se crece el óxido, permaneciendo en el silicio o en el óxido dependiendo del coeficiente de segregación.

11.3 Deposición de dieléctricos

Las láminas de dieléctricos se utilizan principalmente para aislamiento y pasivación de dispositivos discretos y circuitos integrados. Hay tres métodos de deposición:

- CVD a presión atmosférica (Deposición química en fase de vapor)
- LPCVD (Deposición química a bajas presión)
- PCVD (Deposición de plasmas)

Los criterios de selección de uno u otro método son la temperatura del substrato, la velocidad de deposición y uniformidad de la película, la forma del substrato, las propiedades eléctricas y mecánicas, y la composición química de la capa dieléctrica. (Por ejemplo los proceso PCVD se utilizan cuando las dimensiones de los dispositivos alcanzan las regiones submicra debido a que es necesario utilizar bajas temperaturas para evitar la difusión de los dopantes de unas zonas a otras de los dispositivos.

Métodos de deposición de dióxido de silicio

El dióxido de silicio creado mediante un proceso CVD es de peor calidad que el crecido térmicamente. Los óxidos crecidos con CVD se emplean para completar a los óxidos térmicos. Óxidos CVD se utilizan para aislar las capas de metal en metalizaciones de varios niveles, como máscara para la difusión de impurezas e implantación iónica, y para aumentar el espesor de los óxidos crecidos térmicamente. El dióxido de silicio dopado con fósforo se utiliza tanto como aislante entre diferentes láminas de metal y como capa final de pasivación de los dispositivos. Por último, óxidos dopados con fósforo, arsénico o boro se utilizan ocasionalmente como fuentes de difusión.

Las capas de dióxido de silicio pueden ser depositadas por diferentes métodos. A bajas temperaturas (300 a 500 ºC) las películas se forman al reaccionar silano, dopante y oxígeno. Las reacciones químicas para óxidos dopados con fósforo son las siguientes:

$$SiH_4 + O_2 \rightarrow SiO_2 + 2H_2 \qquad (450^{o}C) \qquad (11.19)$$

$$4PH_3 + 5O_2 \rightarrow 2P_2O_5 + 6H_2 \qquad (450^{o}C) \qquad (11.20)$$

La baja temperatura de deposición de la reacción silano-oxígeno hace este proceso adecuado para depositar capas de óxido sobre una capa de aluminio.

En un rango de temperatura intermedias (500 a 800 ºC), el dióxido de silicio puede formarse por descomposición del tetraetilortosilicato $Si(OC_2H_5)_4$ (TEOS) en un reactor LPCVD. Estos óxidos puede doparse añadiendo pequeñas cantidades de hidruros de dopantes (fosfina, arsina, o diborano).

A altas temperaturas (900 ºC) la deposición de dióxido de silicio se consigue haciendo reaccionar diclorosilano, $SiCl_2H_2$ con óxido nitroso a bajas presiones, de acuerdo con la reacción química:

$$SiCl_2H_2 + 2N_2O \rightarrow SiO_2 + 2N_2 + 2HCl \qquad (11.21)$$

Este método proporciona capas de muy uniformes y se usa alguna veces para obtener capa de aislantes sobre polisilicio.

Las propiedades y calidad de las capas de óxido depositadas depende de la temperatura del proceso. En general, a medida que mayor es la temperatura mejor es la calidad del óxido, de manera que a altas temperaturas el óxido depositado se parece al óxido crecido térmicamente. Por el contrario, los óxidos obtenidos a bajas temperaturas son porosos lo que implica una baja calidad dieléctrica, de manera que cuando se aplica un campo eléctrico al óxido se establece a través de él un flujo importante de corriente.

Cobertura de escalones (Step Coverage)

Este concepto está relacionado con la topografía de la capa depositada sobre las diferentes escalones del substrato semiconductor. Cuando la cobertura es homogénea, el óxido crecido tiene el mismo espesor independientemente de los huecos o escalones que presente el substrato. La figura 11.3.1 expone claramente este concepto.

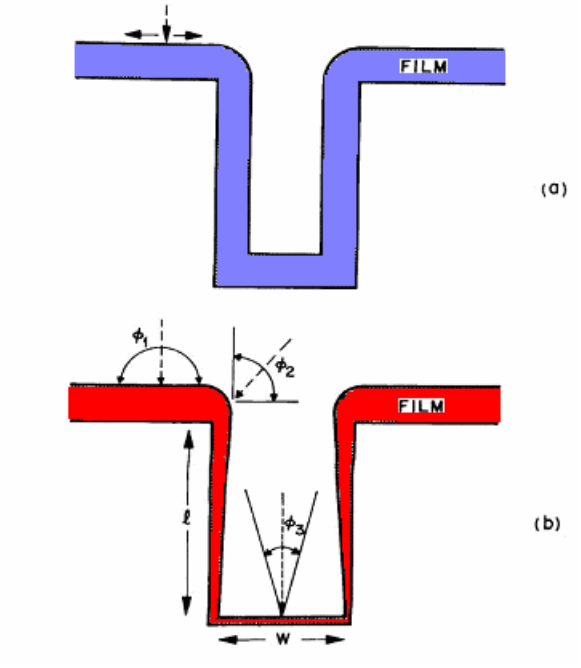


Figura 11.3.1 Cobertura de escalones

En la figura (a), el espesor de la capa depositada es el mismo en todas las zonas del substrato. Es lo que se denomina *conformal step*

coverage. Esta uniformidad se debe a la rápida migración de los reactivos después de la adsorción sobre toda la superficie del substrato.
En la figura (b), el espesor de la capa depositada depende de la zona del substrato. Es lo que se denomina *non conformal step coverage.* En este caso los reactivos no se distribuyen por la superficie del substrato después de la adsorción. En este caso, la velocidad de crecimiento de la capa es proporcional al ángulo de llegada de las moléculas de gas. Por ejemplo, para la superficie superior, las moléculas llegan con ángulos desde 0º a 180º. Para las zonas de las esquinas, sólo llegan aquellas moléculas que tengan un ángulo de incidencia entre 0º y 90º. Por último, a la zona en el interior del pozo, sólo llegan aquellas moléculas con un ángulo de incidencia

$$\phi_3 = arctg\frac{W}{l} \qquad (11.22)$$

donde l es la profundidad del pozo y W la anchura.

El dióxido de silicio obtenido a partir de TEOS proporciona una cobertura uniforme. Sin embargo, la deposición a partir de la reacción del silano con oxígeno proporciona una cobertura que depende del ángulo de llegada de las moléculas.

Deposición de nitruro de silicio

Las capas de nitruro de silicio se utilizan para pasivación de dispositivos, puesto que actúan como barrera a la difusión de agua y sodio.

Pueden usarse también cómo máscaras para la oxidación selectiva del silicio (LOCOS - LOCal Oxidation of Silicon), puesto que el nitruro de silicio se oxida muy lentamente evitando que el substrato de silicio que existe por debajo de él se oxide. La figura 11.3.2 muestra esquemáticamente un proceso de oxidación selectiva mediante el uso de Si_3N_4

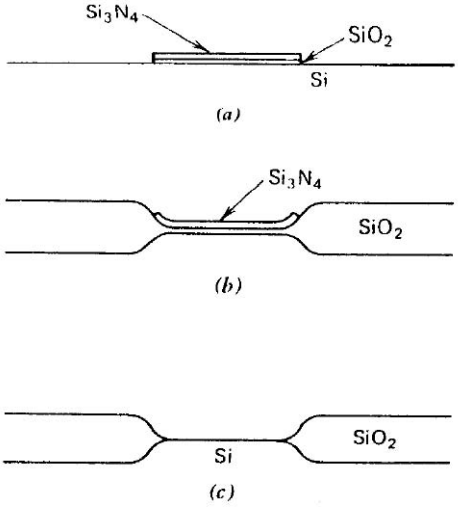


Figura 11.3.2. Técnica de oxidación local de silicio (LOCOS) mediante el uso de Si_3N_4.

Para depositar una capa delgada de nitruro de silicio se hace reaccionar diclorosilano y amoníaco a baja presión y a una temperatura entre 700 y 800º C:

$$3SiCl_2H_2 + 4NH_3 \rightarrow Si_3N_4 + 6HCl + 6H \qquad (11.23)$$

11.4 Deposición de silicio policristalino

El uso del silicio policristalino (o polisilicio) en circuitos integrados MOS ha supuesto un avance tecnológico muy importante. La razón principal es que proporciona electrodos mucho más fiables que el aluminio. La figura 11.4.1 muestra el tiempo máximo de ruptura del electrodo para capacidades MOS que utilizan tanto electrodos metálicos como de polisicilio. Como se observa el polisilicio es claramente superior especialmente para óxidos más delgados. El polisilicio también se utiliza como fuente de difusión para crear uniones poco profundas y para asegurar contacto óhmico con silicio cristalino.

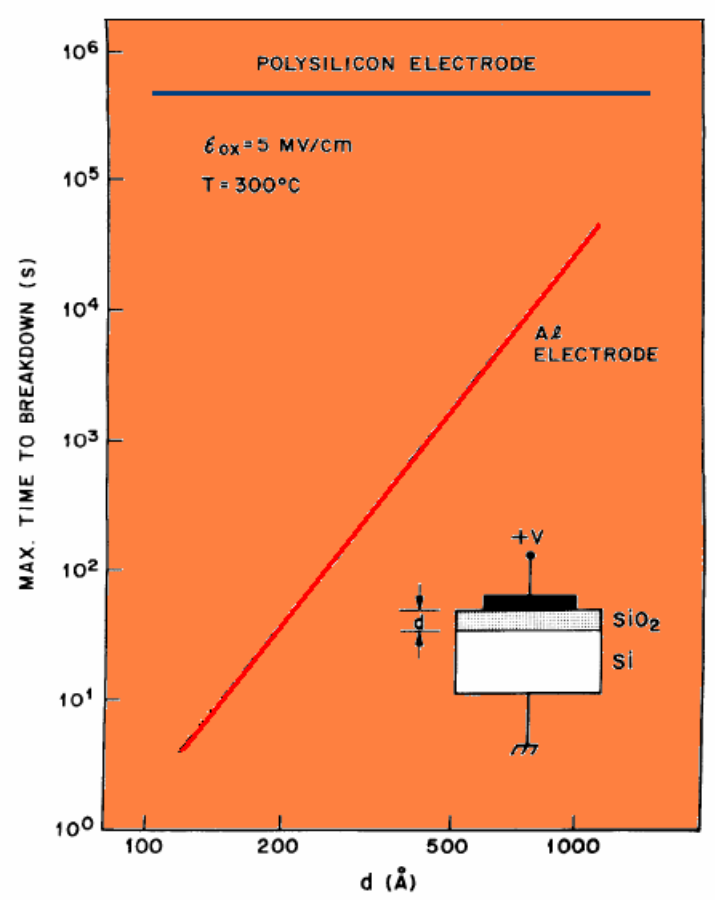


Figura 11.4.1. Tiempo máximo de ruptura del electrodo para capacidades MOS con electrodos metálicos (rojo) o electrodos de polisilicio (azul).

Para depositar polisilicio, se utiliza un reactor LPCVD a una temperatura entre 600 y 650º C donde se produce la pirólisis del silano:

$$SiH_4 \rightarrow Si + 2H_2 \qquad (11.24)$$

La velocidad de crecimiento y la estructura del polisicio depositado, depende de la temperatura y presión parcial del silano empleado, y de los dopantes. El polisilicio puede doparse por difusión, implantación iónica o mediante la adición de gases dopantes durante la deposición. Generalmente se utiliza la implantación iónica debido a la menor temperatura del proceso.

El empleo del polisilicio ha sido especialmente importante en tecnología MOS. Recientemente, sin embargo, ha sido utilizado también en circuitos integrados bipolares.

11.5 Metalización

El proceso de metalización se refiere a la formación de películas metálicas usadas para interconectar los diferentes dispositivos de un mismo circuito integrado, y crear contactos óhmicos o contactos rectificadores metal-semiconductor. Las películas metálicas puede crearse mediante diferentes técnicas, de entre las que destacan PVD (Physical Vapor Deposition) y CVD (Chemical Vapor Deposition).

Physical Vapor Deposition (PVD)

Este proceso se realiza en un ambiente de alto vacio, empleándose técnicas de evaporación o de sputtering (Bombardeo iónico). Los átomos de metal vaporizado por alguna de estas técnicas viajan hacia las obleas donde se condensan para formar una película uniforme. El metal suele ser normalmente aluminio o una aleación de éste (figura 11.5.1).

Para el proceso de evaporación se utiliza alguna de las fuentes siguientes:

En la figura 11.5.2(a) se utiliza un filamento de tungsteno (temperatura de fusión muy elevada). De cada espira del filamento se cuelga un pequeño trozo de aluminio. Este método es simple y barato y no produce radiación ionizante. Presenta sin embargo como desventajas la posible contaminación debida al calentador y a que sólo pueden formarse capas delgadas debido a la pequeña carga de aluminio.

La figura 11.5.2(b) muestra otro posible montaje para conseguir la evaporación del metal. En un crisol de nitruro de boro se deposita el metal empleada para la metalización. El crisol se calienta mediante inducción RF.

La figura 11.5.2(c) muestra el procedimiento de evaporación por haces de electrones (e-beam). Un filamento termoiónico suministra el haz de electrones que son acelerados por un campo eléctrico y conducidos hacia la superficie del metal donde al chocar con éste producen la evaporación del mismo. Si se utilizan diferentes depósitos con metales diferentes pueden depositarse diferentes aleaciones. El principal inconveniente de este proceso es la generación de rayos-X (radiación ionizante que produce la creación de trampas en el óxido y degrada las características eléctricas de los dispositivos).

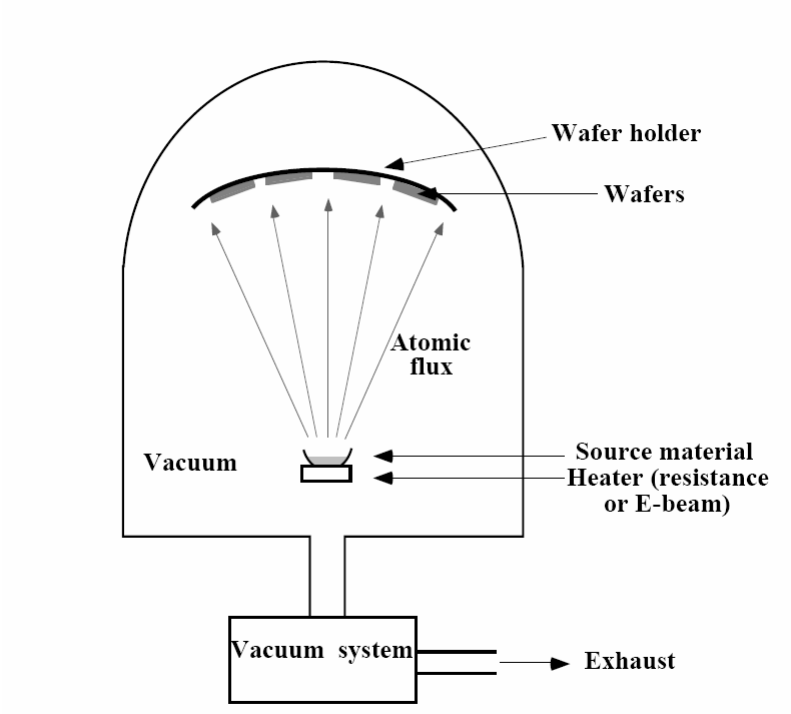


Figura 11.5.1. Sistemas de deposición física en fase de vapor (PVD)

La otra técnica empleada en PVD es el sputtering o bombardeo de iones. El proceso consiste en la aceleración de iones (generalmente Ar^+) a través de un gradiente de potencial y el posterior bombardeo de estos iones sobre un material diana u cátodo (Figura 11.5.3) . Debido a la transferencia de momento desde los iones a los átomos de la superficie del metal diana, estos últimos se convierten en volátiles y son transportados como vapor hacia la superficie del substrato (Figura 11.5.4)

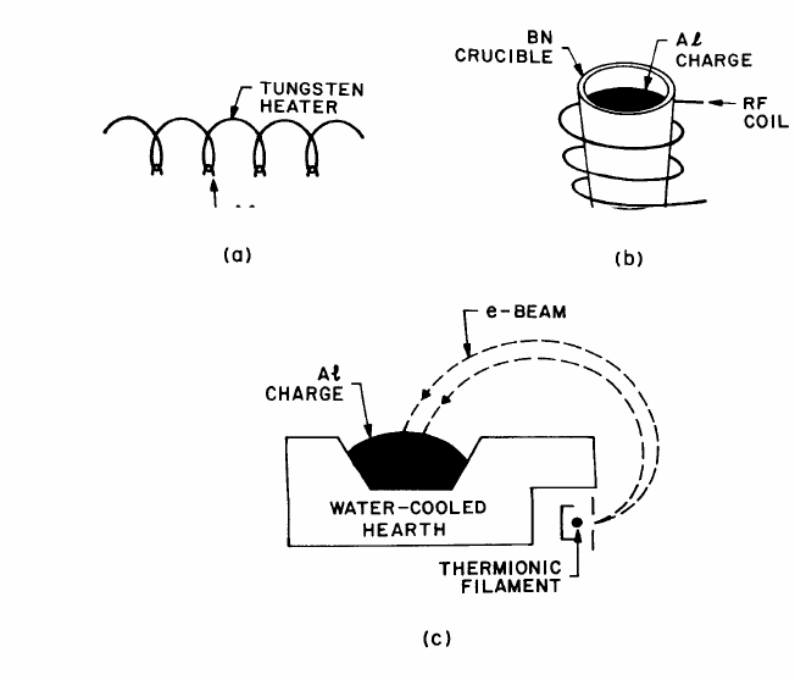


Figura 11.5.2. Distintos sistemas de evaporación de metales.

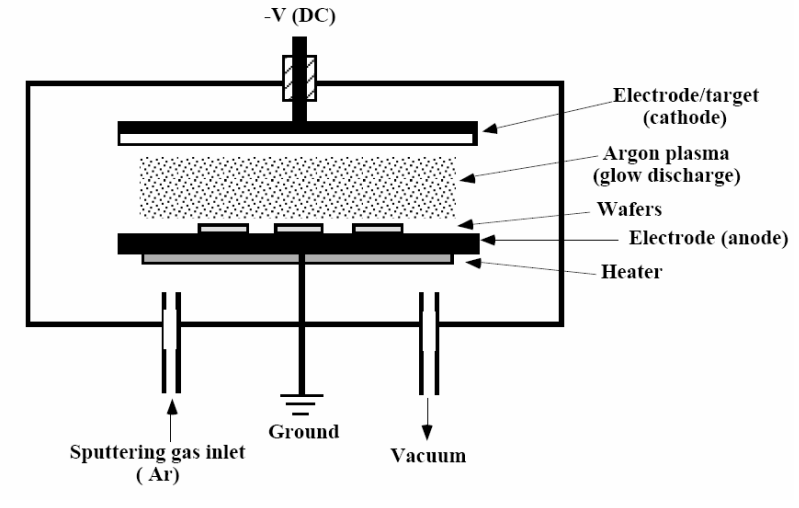


Figura 11.5.3. Diagrama de un sistema de sputtering para metalización.

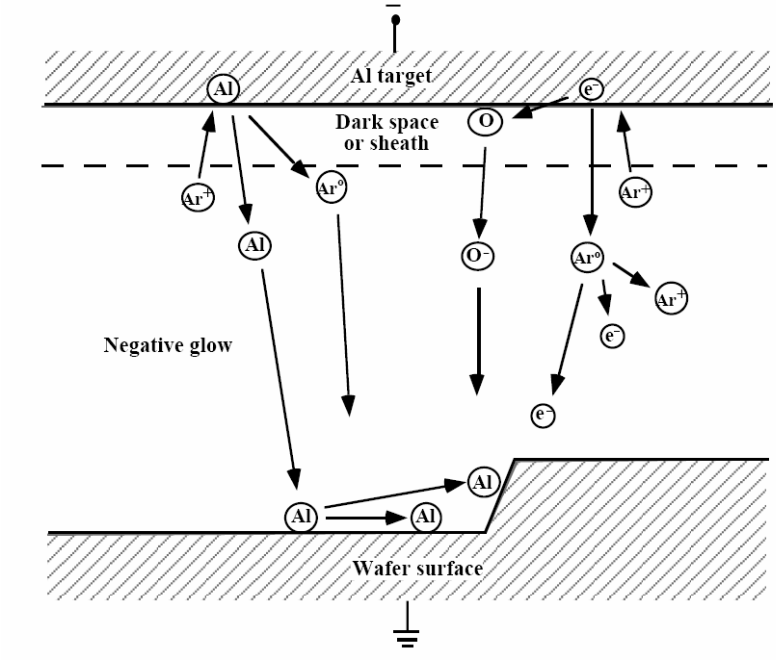


Figura 11.5.4. Esquema del funcionamiento de la deposición de metales por sputtering.

Deposición química en fase de vapor

El otro procedimiento para la deposición de capas metálicas es CVD. Una de las mayores aplicaciones de la deposición de metales con CVD, es la deposición de metales refractarios como el tungsteno, puesto que su elevado punto de fusión y baja resistividad lo hacen un material apropiado para la fabricación de circuitos integrados.

Cuando las dimensiones de los dispositivos se reducen, los requerimientos sobre la metalización son cada vez más severos. La reducción de las dimensiones implican el aumento de la densidad de corriente a través de las conexiones entre los diferentes dispositivos: si las dimensiones del dispositivo se reducen en un factor k, la corriente escalada debería disminuir también en un factor k. Sin embargo, la sección de la interconexión se reduce en un factor k^2, por lo que la densidad de corriente aumenta en un factor k. Este aumento de la densidad de corriente supone un aumento de la tensión que caen entre los extremos de las conexiones. Para minimizar este efecto se requiere. lógicamente, reducir la resistividad de los materiales empleados para la interconexión. Antes hemos visto las ventajas del polisilicio frente al aluminio. Sin embargo, la resistividad del polisilicio es de 500μΩ/cm mientras que la del aluminio es de 2.7 μΩ /cm.. Por esta razón se utilicen siliciuros de metales refractarios ($TaSi_s$, $TiSi_s$) que presentan propiedades

parecidas a las del polisicilio pero con una menor resistividad (50 μΩ/cm).

Electromigración

Un fenómeno relacionado con la interconexión de los circuitos integrados y que supone un serio problema de fiabilidad de los circuitos integrados es la electromigración que puede provocar el mal-funcionamiento de un circuito integrado tras cientos de horas de buen funcionamiento, debido a la ruptura de alguna de sus conexiones internas.

El fenómeno de la electromigración se debe al movimiento de los átomos del material conductor que forma la conexión debida a la transferencia de momento entre los portadores móviles y los átomos del metal. En aluminio, los electrones al moverse colisionan con los átomos que empujan hacia el electrodo positivo. Este trasiego de material hacia el terminal positivo debilita la capa metálica que acaba rompiéndose: En el caso del aluminio la electromigración empieza a ser importante para densidades de corriente del orden de 10^5 A cm^{-2}. Este fenómeno puede reducirse añadiendo pequeñas cantidades de otro metal como cobre. Pueden emplearse también metales refractarios como el tungsteno.

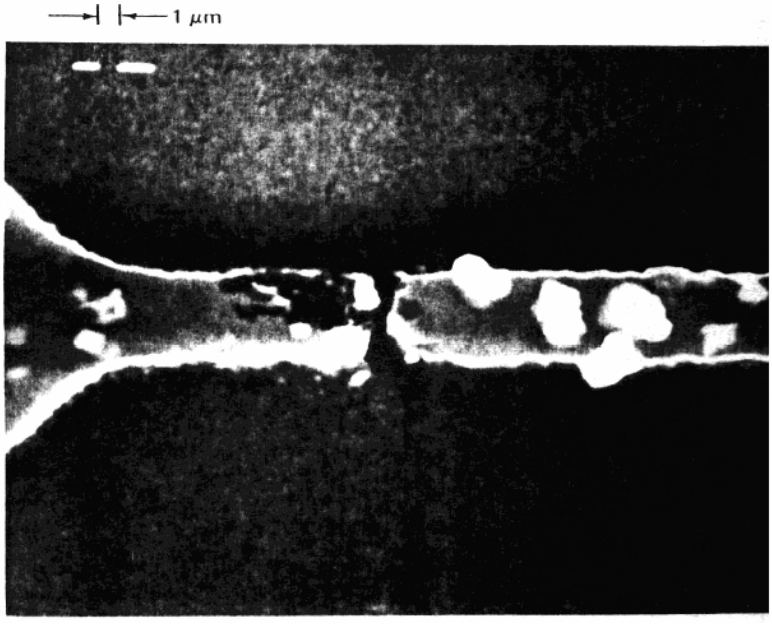


Figura 11.5.5. Fotografía mostrando la ruptura de una conexión metálica por el efecto de la electroemigración.

RESUMEN

En este capitulo se han analizado las técnicas principales para la obtención de diferentes películas delgadas necesarias para la fabricación de dispositivos electrónicos. Estas películas pueden formar parte de los propios dispositivos electrónicos, o son necesarias en el proceso de fabricación de las mismas. Empezamos estudiando la obtención de dieléctricos y principalmente el dióxido de silicio, un buen aislante fácilmente obtenible. Se distingue entre óxidos nativos (aquellos obtenidos directamente del propio silicio del substrato por oxidación) y de los óxidos depositados (aquellos en los que el óxido se obtiene a partir de una fuente de silicio externa). Se estudian las diferentes propiedades y usos del dioxido de silicio dependiendo de la técnica utilizada para su obtención. A continuación se estudia la deposición de otros dieléctricos, como el nitruro de silicio. Finalmente se estudia la deposición de películas de polisilicio y películas metálicas.

CUESTIONES Y PROBLEMAS

1. Se oxida una muestra de silicio en ambiente húmedo a una temperatura de 900ºC durante 15 minutos. Calcule el espesor de óxido formado. Una vez finalizado el proceso anterior se prosigue con una oxidación en ambiente seco a la misma temperatura durante 30 minutos. Obtenga el espesor final del óxido.

2. Obtenga el tiempo necesario para obtener un espesor de óxido de 0.5µm en una oxidación seca. Repita el apartado anterior suponiendo que la oxidación es una oxidación húmeda. En ambos casos la temperatura del proceso es de 1000ºC.

REFERENCIAS

[1] S. Sze. *VLSI Technology*. Ed. McGraw-Hill.

[2] Chang and S. Sze. *ULSI Technology*. Ed. McGraw-Hill.

[3] J.D.Plummer, M.D.Deal, P.B.Griffin, *Silicon VLSI Technology*, Ed.Prentice Hall.

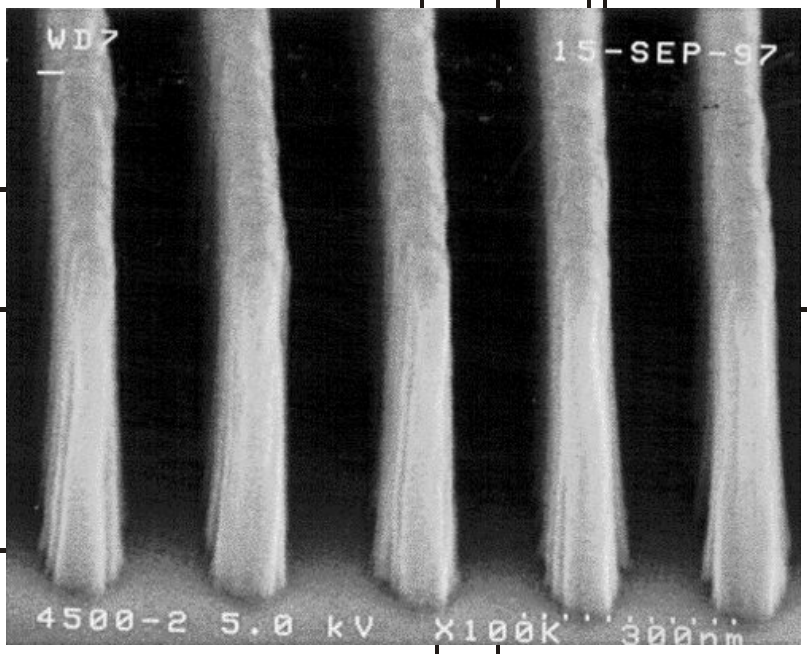
[4] Streetman and Banerjee. *Solid State Electronic Devices*. Prentice Hall, Fifth edition, 2000.

[5] Glosario: http://semiconductorglossary.com/

12

Capítulo

Litografía y Grabado



Índice

Objetivos

- Describir los conceptos de litografía y grabado.
- Explicar los diferentes tipos de procesos existentes.
- Comentar los problemas y soluciones que se presentan a medida que las dimensiones de los dispositivos, que se quieren fabricar, son mas pequeñas.
- Explicar como son las instalaciones donde se realizan estos procesos, las llamadas Salas Blancas y que condiciones deben cumplir.

Palabras clave

Litografía UV
Litografia rayos X
Haces de electrones
Fotorresina positiva
Fotorresina negativa
Mascara
Grabado seco
Grabado humedo
Ataque isotrópico
Ataque anisotrópico
Selectividad en el grabado
Sala Blanca
Escalonador
Reactive Ion Etching (RIE)

12.1 Litografía

Una vez creada la capa de aislante SiO2 sobre la oblea, parte de ella debe ser eliminada selectivamente en aquellos sitios en los que den introducirse los átomos de dopante. El grabado selectivo se realiza generalmente mediante el uso de un material sensible a la luz denominado **fotorresina**. Para ello, la oblea oxidada se cubre en primer lugar por una capa de fotorresistencia. A continuación se recubre la fotorresistencia con un negativo fotográfico parcialmente transparente denominado máscara o fotomáscara.

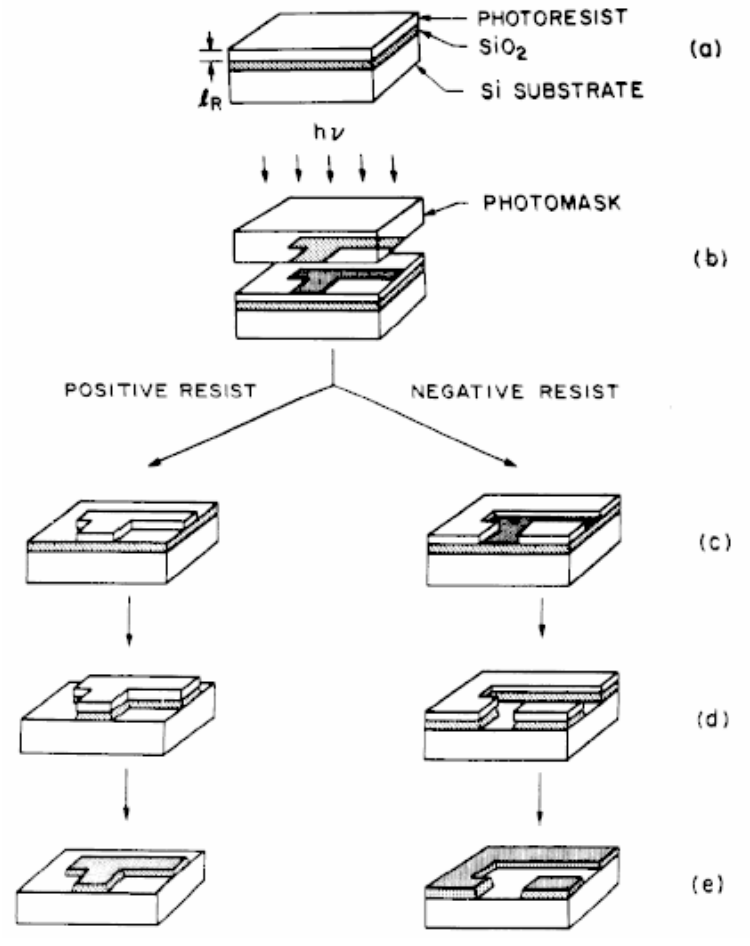


Figura 12.1.1: Esquema de un proceso de litografía óptica.

Una vez convenientemente alineada la máscara se ilumina con luz ultravioleta. La luz ultravioleta cambia la estructura de la fotorresistencia: las moléculas de una **fotorresistencia negativa** se unen entre si (polimerizan) en las regiones expuestas a la luz. Por el contrario, en el caso de **fotorresistencias positivas,** los enlaces entre las moléculas se rompen al iluminarse, permaneciendo polimerizadas el resto. Las partes no iluminadas de las fotorresistencias no se ven afectadas. Las áreas no

polimerizadas de la fotorresistencia (no iluminadas en el caso de fotorresistencia negativa e iluminadas en el caso de fotorresistencias positivas) se disuelven selectivamente usando por ejemplo tricloroetileno. De esta forma las zonas polimerizadas, resistentes al ataque del ácido quedan protegiendo al SiO_2 .

Sala blanca

En el proceso de litografía debe realizarse en una habitación espacial denominada sala blanca. La necesidad de esta sala blanca se debe a que las partículas de polvo del aire pueden depositarse sobre la superficie del semiconductor y las máscaras litográficas causando defectos en los dispositivos. Cuando las partículas de polvo se depositan sobre las fotomáscaras, se comportan como cuerpos opacos de la misma forma que los diferentes patrones grabados sobre la máscara (las partículas de polvo tienen dimensiones del orden de los diferentes patrones). La figura 12.1.2, muestra los posibles efectos de una partícula de polvo depositada sobre una fotomáscara.

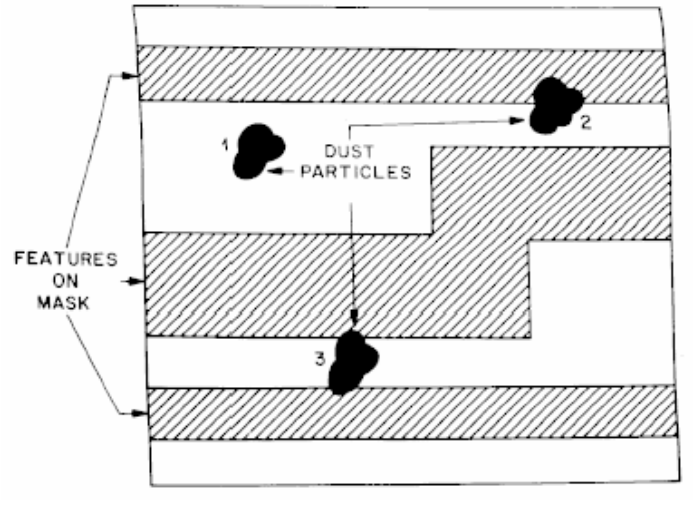


Figura 12.1.2: Efectos de una partícula de polvo sobre una máscara litográfica.

La partícula 3 da lugar a un cortocircuito entre las dos regiones conductoras.

En una sala blanca, se controla con precisión el número total de partículas de polvo por unidad de volumen. La figura 13.1.3 muestra la distribución del tamaño de partículas para diferentes clases de salas blancas.Así por ejemplo, una sala blanca clase A, tiene 100 partículas con diámetro 0.5:m o mayor por pie cúbico (corresponde a 3500 partículas por metro cúbico). Se observa que el número de partículas permitido aumenta a medida que disminuye el diámetro de estas. Para el proceso litográfico se requiere normalmente una sala blanca clase 10.

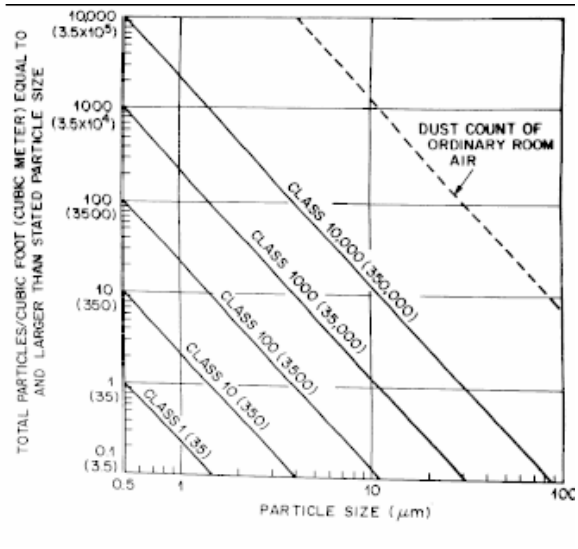


Figura 12.1.3.: Distribución de particulas según el tipo de clase.

Si la dimensión mínima de los dispositivos se aproxima a la longitud de onda de la luz utilizada en la exposición óptica para sensibilizar la fotoresina ($\lambda \sim 400$ nm) los fenómenos de difracción pueden limitar la resolución del método (tamaño mínimo que puede distinguirse). Para evitar esta limitación se han propuesto técnicas alternativas. Un procedimiento alternativo a la fotolitografía con luz convencional, es utilizar luz ultravioleta, que al tener menor longitud de onda reduce los fenómenos de difracción. Las ventajas de este simple método son limitadas de ahí que se consideren otro tipo de técnicas, como el uso de haces de electrones, rayos x, o haces de iones para sensibilizar la fotorresistencia.

-Litografía por haces electrónicos: Un chorro de electrones energéticos se dirige sobre la fotorresistencia que queda sensibilizada. En vez de sensibilizar todos los patrones a la vez, se van "dibujando" uno a uno los distintas partes del circuito integrado, por lo que no es necesaria ninguna máscara. Aunque se consigue de esta forma una resolución mucho mayor que cualquier dimensión del circuito integrado, presenta el inconveniente de ser un proceso lento, puesto que hay que grabar uno a uno los diferentes partes del circuito integrado.

-Litografía por rayos x: Un haz de rayos X se hace pasar por una máscara para sensibilizar selectivamente la fotorresistencia. Al igual que la fotolitografía convencional, la litografía con rayos X permite grabar varios patrones de forma simultánea. Además, puesto que la longitud de onda es mucho más pequeña, se consigue también de esta forma una mejor resolución, y por lo tanto unos dispositivos de menor tamaño. Los inconvenientes fundamentales de este tipo de fotolitografía es que las máscaras son difíciles de fabricar y que además la utilización de rayos X puede dañar las partes activas de los dispositivos.

12.2 Grabado

Una vez que los patrones se han grabado sobre la fotorresistencia por alguna de las técnicas de litografía estudiadas, y se ha disuelto la parte no polimerizada de ésta mediante el uso de tricloroetileno, es necesario eliminar la parte de SiO_2 no protegida para abrir las ventanas deseadas en el óxido, que dejen a la vista el substrato de silicio. Para eliminar la parte de óxido no protegido puede usarse un baño de ácido fluorhídrico que ataca al dióxido de silicio no protegido, pero no ataca al silicio. Existen un gran número de diferentes reactivos químicos que atacan a los materiales selectivamente de manera que eliminan la capa de óxido pero producen un ataque muy pequeño sobre los materiales subyacentes. Esta técnica se denomina **grabado químico o grabado húmedo.**

Generalmente el ataque químico de la capa de óxido suele ser isotrópico, es decir, por igual en todas las direcciones (aunque existen algunas excepciones). Esto significa que el óxido no sólo se ataca hacia abajo sin que también se ataca lateralmente por debajo del protector:

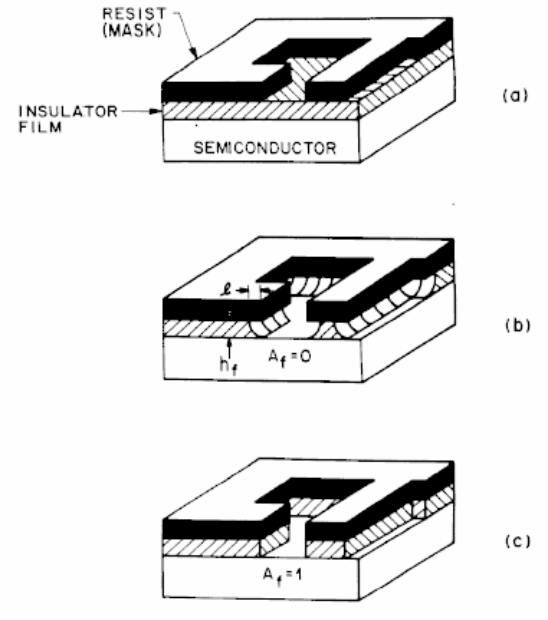


Figura 12.2.1. Comparación de ataque isotrópico (b) y anisotrópico (c)

En consecuencia las ventanas abiertas en el óxido, son por este motivo mayores que las marcadas en máscara. Esto que puede llegar a ser un inconveniente importante cuando se reducen las dimensiones de los dispositivos. Esta dificultad se resuelve satisfactoriamente con los **grabados en seco-** (Dry etching) en un proceso de grabado en seco, la oblea se expone a un plasma, que consiste en un gas parcial o totalmente ionizado compuesto de iones, electrones y neutrones. El plasma se produce cuando un campo eléctrico de suficiente magnitud se aplica al gas, causando la ionización de las moléculas o átomos del gas. El plasma

se inicia por electrones libres que tras ser proporcionados por un electrodo polarizado negativamente adquieren energía cinética gracias al campo eléctrico. En el transcurso de su viaje a través del gas los electrones chocan con las moléculas del gas y pierden su energía. La energía transferida durante las colisiones hace que las moléculas del gas se ionicen. Las concentración de electrones en un plasma empleado para grabado en seco suele ser baja, del orden de 10^9 a 10^{12} cm^{-3}. Las moléculas ionizadas del plasma son aceleradas perpendicularmente a la oblea, donde chocan con los átomos del semiconductor, que adquieren energía para liberarse de los enlaces que los mantienen unidos al mismo. (en otros casos los moléculas ionizadas reaccionan químicamente con el material que debe ser atacado, con lo que se consigue una mayor selectividad (RIE.reactive-ion-etching). De esta forma se eliminan los átomos de la superficie, llevándose a cabo el grabado. Como no se produce bombardeo sobre las paredes laterales el grabado lateral es mucho más pequeño que el grabado perpendicular, con lo que se consigue un alto grado de anisotropía. Existe sin embargo un inconveniente importante, que es la pérdida de selectividad en el grabado, de tal forma que también el substrato se ve afectado por el ataque.
El proceso de acción de grabado por plasma queda esquematizado en la Figura 12.2.2

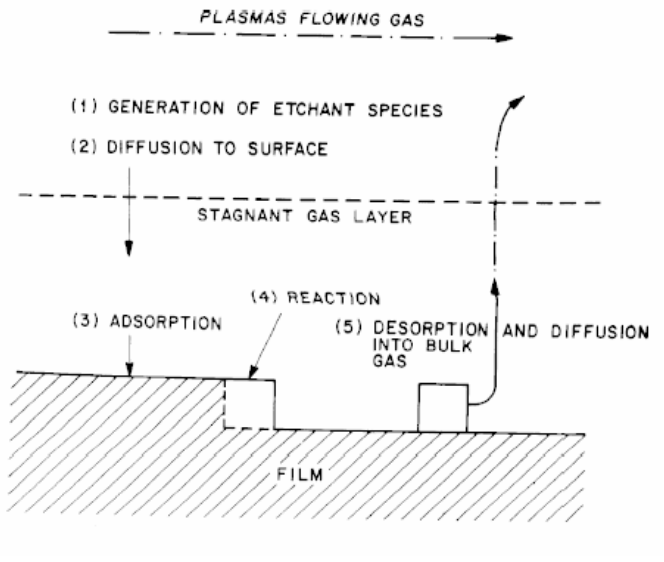
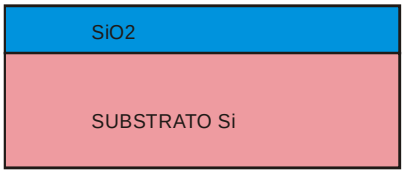


Figura 12.2.2 Esquema de funcionamiento del ataque RIE

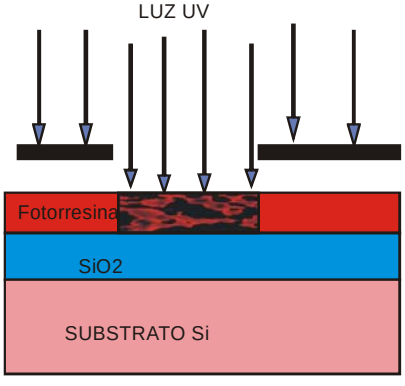
(1) El proceso comienza con la formación de los reactivos
(2) Los reactivos son transportados por difusión a través de una capa gaseosa de estaño hacia la superficie.
(3) La superficie adsorbe a los reactivos.
(4) Se produce la reacción química de los reactivos con la especies de la superficie, junto con efectos físicos (bombardeo iónico).
(5) Los materiales resultados de la reacción química o bombardeo físico son repelidos por la superficie y eliminados por un sistema de vacío.

Proceso de Litografía y Grabado usando una fotorresina positiva.

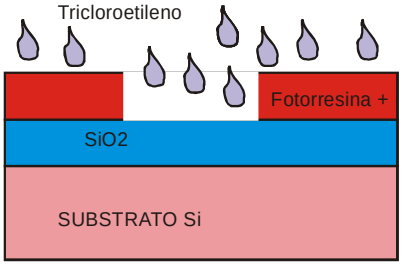


(a) Oblea despues de haber realizado el proceso de oxidación termica

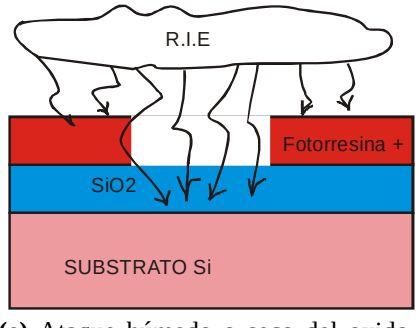
(b) Recubrimiento con una resina fotorresistente positiva



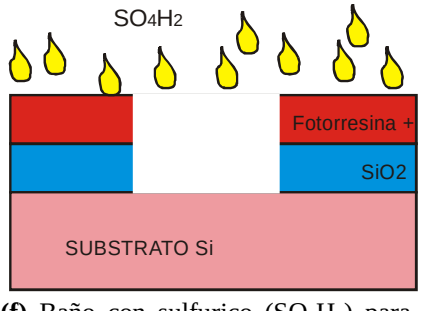
(c) Colocación de la mascara alineada y exposición a fuente de luz.



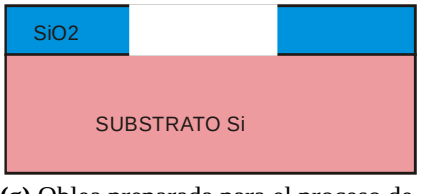
(d) Tratamiento con un disolvente (tricloroetileno) para eliminar la resina expuesta a la luz.



(e) Ataque húmedo o seco del oxido no protegido por la resina.



(f) Baño con sulfurico (SO_4H_2) para eliminar la capa de resina



(g) Oblea preparada para el proceso de implantación

REFERENCIAS

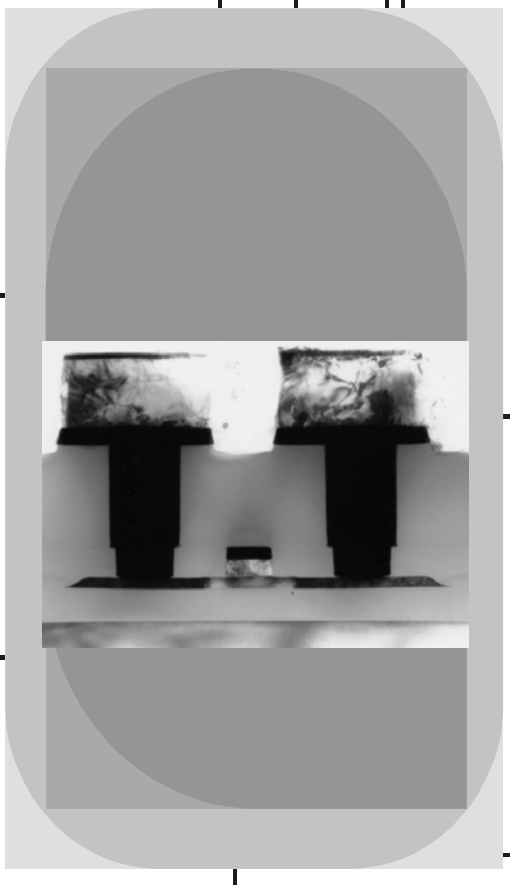
[1] S. Sze. *VLSI Technology*, Ed. Mcgraw-Hill

[2] J.D.Plummer, M.D.Deal, P.B.Griffin, *Silicon VLSI Technology*, Ed.Prentice Hall

13

Capítulo

Tecnología de fabricación de circuitos



SOI MOSFET

Índice

Objetivos

- Descripción de la secuencia de fabricación de componentes pasivos en tecnología bipolar y MOS
- Descripción de la secuencia de fabricación de transistores bipolares
- Descripción de la secuencia de fabricación de transistores NMOS, CMOS y BiCMOS
- Descripción de la secuencia de fabricación de obleas de silicio sobre aislante (SOI)
- Descripción de la fabricación de transistores MESFET en Arseniuro de Galio

Palabras Clave

Resistencia laminar
Capacidades integradas
NMOS, CMOS, SOI
BiCMOS, MESFET
Difusión
Implantación
Oxidación
Si_3N_4
Máscara
Fotorresistencia
Litografía
Grabado
Eltran, Smart Cut
Silicio Poroso
Óxido enterrado
Aislamiento en trinchera y meseta
Latchup

13.1 Introducción

La mayoría de los sistemas electrónicos se implementan utilizando circuitos integrados. Un circuito integrado es un conjunto de dispositivos activos y pasivos construidos sobre un mismo substrato semiconductor e interconectados mediante diferentes capas de metalización. Los elementos más importantes de un circuito integrado son los transistores, por lo tanto, describiremos cómo se fabrican los transistores en las diferentes tecnologías existentes. La figura siguiente muestra los principales pasos a seguir en la fabricación de un circuito integrado.

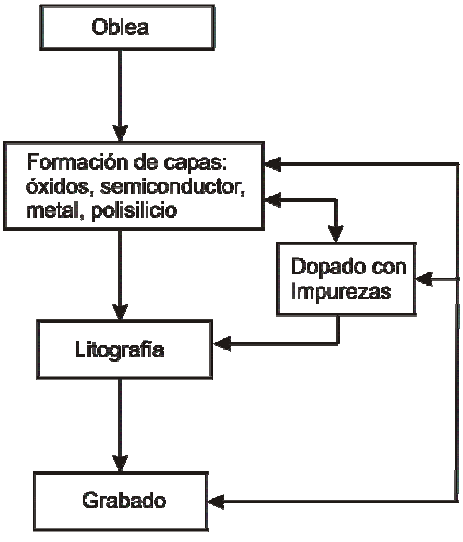


Figura 13.1.1 Diagrama de flujo en el proceso de fabricación de circuitos integrados.

A la hora de diseñar un circuito integrado es necesario tener en cuenta las siguientes reglas que son diferentes de las de diseño de un circuito discreto:

- El material de partida en la construcción de un circuito integrado es una oblea semiconductora con una determinada resistividad y orientación cristalográfica. El primer proceso consiste generalmente en la formación de capas delgadas de material sobre la oblea. Esta se puede realizar mediante procesos de crecimiento epitaxial de películas semiconductoras, crecimiento térmico de óxidos, y deposición de polisilicio de capas dieléctricas y metálicas.
- Tras la formación de las diferentes películas de material, el siguiente paso suele ser el de litografía o el de introducción de impurezas (bien por difusión o por implantación iónica).

- El proceso litográfico normalmente va seguido por un grabado (etching) al que pueden seguir nuevamente otros procesos de dopado y/o formación de películas delgadas.

Después de realizar la secuencia apropiada, cada oblea contiene cientos de chips rectangulares idénticos de 1 a 10mm de lado. Cada chip es comprobado eléctricamente y los defectuosos se marcan con una mancha de tinta negra. A continuación, los diferentes chips son cortados y separados. Aquellos que han superado satisfactoriamente todos los tests, se encapsulan. De esta manera se consigue un buen aislamiento térmico y eléctrico y un entorno adecuado para la utilización del circuito integrado en diferentes aplicaciones electrónicas.

Un chip puede contener desde unos cuantos dispositivos (activos o pasivos) hasta varios millones. Desde la invención del circuito integrado en 1958, el número de componentes ha crecido exponencialmente.

Generalmente se hace la siguiente clasificación de los circuitos integrados:

-SSI (small scale of integration) hasta 10^2 componentes.

-MSI (medium scale of integration) hasta 10^3 componentes.

-LSI (large scale of integration) hasta 10^4 componentes.

-VLSI/ ULSI (very/ultra large scale of integration) más de 10^5 componentes.

13.2 Componentes pasivos en un Circuito Integrado

Resistencias en un circuito integrado

Para construir una resistencia en un circuito integrado se define una ventana en una capa de óxido térmicamente crecido sobre un substrato de silicio. A continuación se implantan (o se difunden) impurezas del tipo contrario a las ya existentes en la oblea.

Si se considera una resistencia lineal, la conductancia de una lámina delgada de material tipo *p* y espesor *x* paralela a la superficie viene dada por

$$\begin{aligned} J &= \sigma E = q\mu_p p(x)\frac{V}{L}; \quad A_s = x \cdot W \\ G &= \frac{1}{R} = \frac{I}{V} = \frac{J \cdot A_s}{V} = q\mu_p p(x)\frac{W}{L}x\ . \end{aligned} \tag{13.1}$$

Según esta expresión podemos escribir la conductancia diferencial como

$$dG = q\mu_p p(x)\frac{W}{L}dx\ , \tag{13.2}$$

donde *W* es la anchura de la barra y *L* su longitud. La conductancia total de toda la región implantada se calcula como

$$G = \int_0^{x_j} dG = q\frac{W}{L}\int_0^{x_j} \mu_p p(x)\,dx \ . \qquad (13.3)$$

Se define la resistencia laminar, $R_\square$:

$$R_\square = \frac{1}{q\int_0^{x_j} \mu_p p(x)\,dx} \ , \qquad (13.4)$$

que se mide en ohmios por cuadrado $(\Omega/\square)$. Por lo tanto la resistencia de la barra se calcula como

$$R = \frac{L}{W} R_\square \ . \qquad (13.5)$$

En consecuencia, el valor de la resistencia en un circuito integrado depende por un lado de los parámetros geométricos *W* y *L*, y por otro lado del valor de la resistencia laminar, $R_\square$, que depende a su vez del proceso de implantación. Una vez que $R_\square$ es conocida, el valor de la resistencia viene dada por la relación L/W, o por el número de cuadrados de dimensiones W×W que contiene el patrón de la resistencia. Los contactos también introducen una resistencia adicional.

Capacidades integradas

Los primeros circuitos integrados se diseñaron pensando que los valores prácticos de las capacidades eran imposibles de integrar debido al gran área que necesitarían y por tanto se hacía uso de capacidades externas. Todavía es cierto que las capacidades integradas con valores superiores a unas decenas de picofaradios son muy costosas en términos de área consumida, sin embargo los diseños han cambiado de tal manera que pequeñas capacidades pueden realizar funciones que antes necesitaban valores muy elevados. Un buen ejemplo es la compensación en los amplificadores operacionales. Ahora se utiliza un gran número de capacidades integradas en casi todos los circuitos integrados. La fabricación de estos componentes es diferente dependiendo de si estamos trabajando en tecnología bipolar o MOS.

En tecnología bipolar encontramos dos tipos diferentes. En primer lugar se utilizó el hecho conocido de que una unión *pn* polarizada en inverso presenta una capacidad de deplexión. No obstante aparecen inconvenientes tales como que la unión debe mantenerse siempre polarizada en inversa, que el valor de la capacidad varía con la tensión aplicada y que para una unión similar a la base-emisor la tensión de ruptura es de sólo 7V. Para la unión base-colector la tensión de ruptura es mayor pero la capacidad por unidad de área es bastante baja.

Por estas razones la capacidad integrada más utilizada en tecnología bipolar es la capacidad MOS.

En la secuencia de fabricación normal se añade un paso adicional en el que se utiliza una máscara para definir una región en la que se crece una delgada lámina de óxido sobre una difusión de emisor y a continuación se realiza una metalización de aluminio sobre el óxido. Queda definida una capacidad entre el aluminio y la difusión de emisor con un valor comprendido entre 0.2 y 0.3 pF/mm^2 y una tensión de ruptura de entre 60 y 100V. Esta capacidad es extremadamente lineal y presenta un coeficiente de temperatura muy bajo. Aparece un inconveniente en forma de capacidad parásita inherente a la región de vaciamiento que se forma entre el substrato tipo *p* y la región epitaxial *n*. No obstante, es despreciable en la mayoría de los casos.

En tecnología MOS las capacidades juegan un papel más importante que en tecnología bipolar ya que éstas desempeñan muchas funciones que en el caso bipolar desarrollan las resistencias.

Diferentes procesos de fabricación MOS utilizan dos láminas de polisilicio para implementar funciones analógicas. La segunda lámina proporciona una estructura capacitiva eficiente y una línea de interconexión extra. La separación entre láminas es comparable al espesor de óxido de puerta de los transistores MOS.

Se debe tener en cuenta la existencia de capacidades parásitas asociadas a cada una de las láminas de polisilicio. La más grande es la que se forma entre la lámina inferior y el substrato, proporcional al área de la lámina inferior y con un valor típico que ronda entre el 10 y el 30 % de la capacidad total. La capacidad parásita asociada a la lámina superior tiene su origen en la metalización que conecta dicha lámina con el resto del circuito más la capacidad parásita del transistor al cual está conectado. El valor de esta capacidad parásita está comprendido entre 5 y 50 fF.

Otros parámetros importantes de estos componentes son la tolerancia, el coeficiente de tensión y el coeficiente de temperatura. La tolerancia en el valor absoluto de la capacidad es una función del espesor del óxido y se mueve en un rango de entre el 10 y el 30%. Sin embargo dentro del mismo chip, las diferencias entre capacidades se reducen a menos de 1%. Esto se debe al hecho de que las láminas que forman los contactos de la capacidad están constituidas por semiconductores muy dopados en lugar de conductores ideales. La realidad es que se producen variaciones en el potencial superficial de la lámina al aplicar la tensión (fenómeno de poli deplexión) resultando en un ligero cambio en la capacidad con la tensión aplicada.

El transistor MOS en sí mismo se puede utilizar también como una capacidad cuando se polariza en la región triodo; la puerta constituye un contacto y la fuente, el drenador y el canal la otra. Debido a que el substrato no está muy dopado se producen grandes variaciones del potencial al modificar la tensión aplicada y por tanto presenta un coeficiente de tensión muy elevado.

13.3 Tecnología Bipolar

Muchos circuitos con aplicación comercial necesitan aumentar el ancho de banda de manera que puedan trabajar a frecuencias más elevadas. Esta necesidad de operar a mayores velocidades se traduce en una reducción de la anchura de la base para así disminuir el tiempo de tránsito de los portadores y el valor de las capacidades parásitas. La reducción de las dimensiones del dispositivo obliga a que la anchura de las regiones de deplexión dentro de la estructura se reduzca en proporción, por lo que es necesario el uso de menores tensiones de operación y mayores concentraciones de impurezas en las distintas regiones que componen el dispositivo. Para cubrir estas necesidades se ha desarrollado una secuencia de procesos para fabricar transistores bipolares diferente a la que se utilizaría, por ejemplo, en aplicaciones de potencia donde es frecuente la aplicación de grandes tensiones. Como diferencias más destacas con respecto a otro tipo de aplicaciones se puede mencionar el uso de láminas epitaxiales más delgadas y dopadas, oxidaciones selectivas en diferentes regiones para conseguir el aislamiento eléctrico entre regiones en lugar de uniones polarizadas en inverso y el uso del polisilicio como fuente de dopantes para el emisor.

Partiendo de un substrato de silicio tipo *p*, se comienza con una primera máscara que permite una implantación, obteniéndose una región n^+ en el substrato *p*. A continuación se lleva a cabo el crecimiento epitaxial de una lámina de silicio tipo *n* con un espesor aproximado de 1micra y alrededor de 0.5 Ω-cm de resistividad. El resultado se muestra en la siguiente figura,

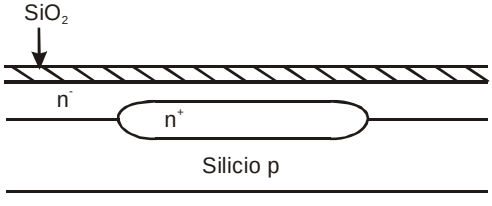


Figura 13.3.1 Sección transversal de la estructura resultante tras la formación de una lámina enterrada n^+ y el crecimiento epitaxial de una lámina tipo *n*.

El siguiente paso consiste en realizar una oxidación selectiva que permite aislar el transistor de sus vecinos y el contacto de colector del resto del transistor. Antes de crecer una gruesa lámina de óxido, se lleva a cabo un grabado que elimina el silicio de aquellas regiones donde se quiere situar el óxido. Sin este paso previo el óxido resultante presentaría un perfil abultado y poco uniforme que dificultaría o impidiría depositar sobre esas zonas láminas de metal o polisilicio. Por tanto, la eliminación del silicio antes de la oxidación permite conseguir una superficie casi plana después de la oxidación y elimina el problema del posterior

recubrimiento con otros materiales. La estructura resultante es la que aparece en la siguiente figura,

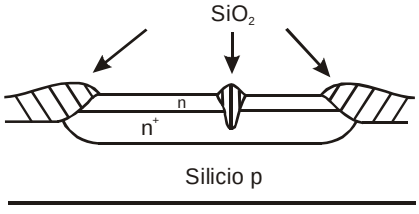


Figura 13.3.2 Sección transversal del dispositivo resultante después de realizar un grabado selectivo y la posterior oxidación.

Las zonas oxidadas se extienden hasta alcanzar el substrato *p* aislando eléctricamente las regiones *n* crecidas epitaxialmente. El crecimiento de láminas de óxido de grosor mayor que 1 micra requiere tiempos muy largos, por este motivo este método de aislamiento es útil exclusivamente para estructuras muy finas.

A continuación se van a definir los contactos de base y colector comenzando con una implantación de impurezas donadoras en elevada concentración en la región de contacto del colector y una difusión posterior en la lámina enterrada, dando lugar a un camino de baja resistencia hasta el colector. Seguidamente se utiliza una segunda máscara para definir la región de base junto con un implante de impurezas aceptadoras. El resultado final se muestra en la figura,

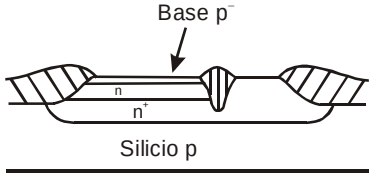


Figura 13.3.3 Sección transversal del dispositivo resultante después de utilizar una máscara para implantar y difundir impurezas donadoras en la región de colector y usar otra máscara e implantar impurezas aceptadoras en la región de base tipo *p*.

Un reto fundamental en la fabricación de este tipo de dispositivos es la formación de una base y emisor muy delgados y al mismo tiempo conseguir un camino de baja resistencia para los portadores hasta los contactos. Normalmente se consigue utilizando polisilicio como fuente de impurezas. Se deposita una lámina de polisilicio n^+ justamente encima del emisor. Durante los posteriores ciclos térmicos a elevadas temperaturas los dopantes (normalmente arsénico) se difunden desde el polisilicio hacia el interior del silicio cristalino formando una región de emisor muy delgada y muy dopada. Siguiendo a la deposición del polisilicio se realiza un implante de boro que da lugar a una lámina de silicio tipo p^+ en todos

los puntos de la base excepto justamente bajo el polisilicio ya que éste actúa como una barrera que impide a los átomos de boro alcanzar el substrato. La estructura que se obtiene tras este paso se observa en la figura,

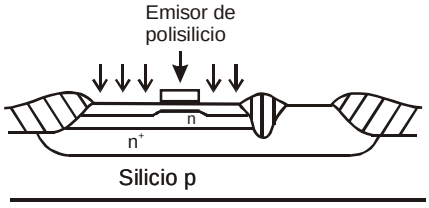


Figura 13.3.4 Sección transversal del dispositivo resultante después de la deposición del polisilicio y usar una máscara para implantar impurezas aceptadoras en la región de base.

El método utilizado para formar contactos de baja resistencia en la base se denomina estructura auto-alineada porque el alineamiento de la región de base con el emisor se produce automáticamente. Un proceso similar se utiliza también en tecnología MOS como se verá en el apartado correspondiente. La estructura que se obtiene después de la metalización se muestra en la siguiente figura,

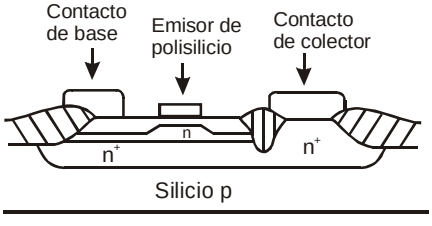


Figura 13.3.5 Sección transversal del dispositivo final resultante. Los contactos de base y colector pueden solaparse con los óxidos permitiendo una reducción de sus dimensiones. El contacto de emisor se realizaría en una extensión del polisilicio mostrado en la figura.

Puesto que las zonas de aislamiento están compuestas de SiO_2, las ventanas de metalización pueden solaparse con ellas. Esto reduce considerablemente las dimensiones mínimas que se pueden conseguir en las regiones de base y colector. Todas las superficies de silicio y polisilicio se cubren con un siliciuro metálico para reducir las resistencias de contacto. En dispositivos de dimensiones reducidas el contacto de emisor se realiza extendiendo el polisilicio fuera del área activa del dispositivo y formando el contacto metálico con el polisilicio allí. No obstante, esta solución añade una resistencia serie de emisor. Los circuitos integrados fabricados con una secuencia de procesos similar a la que acabamos de describir producen transistores bipolares con valores de f_T superiores a 10 GHz que es muy superior a los valores típicos de 500

MHz que se consiguen en los procesos diseñados para soportar elevadas tensiones.

13.4 Tecnología NMOS

En la actualidad los transistores MOS son los más utilizados en circuitos VLSI ya que pueden ser escalados a dimensiones más pequeñas que el resto de dispositivos. La tecnología MOS puede dividirse en tecnología NMOS que produce transistores canal *n*, y tecnología CMOS compuesta por transistores canal *n* y canal *p* sobre el mismo substrato. Las dos tecnologías son importantes puesto que la tecnología NMOS es más simple que la tecnología bipolar (involucra menos pasos de proceso) mientras que la tecnología CMOS proporciona circuitos con muy bajo consumo de potencia. A principio de los 70, la longitud mínima de los transistores MOS era del orden de 7.5 μm y el área total del chip del orden de 6000 μm^2. En la actualidad se fabrican transistores con longitud de canal por debajo de 0.1 μm (100nm).

Proceso de fabricación

El material inicial para la fabricación de transistores NMOS es una oblea de silicio tipo *p*, ligeramente dopada (~ 10^{15} cm^{-3}) y orientada, generalmente en la dirección <100>, ya que posee menor densidad de trampas en las interfases que la <111>, lo que proporciona mejores propiedades eléctricas a los dispositivos (mayor movilidad). Se pueden destacar los siguientes puntos:

1.- El primer paso en la fabricación del transistor NMOS es la formación del óxido de aislamiento. El proceso es parecido al utilizado en tecnología bipolar para crecer el óxido que aísla lateralmente a los distintos dispositivos. Una capa de óxido delgado (~500 Å) se crece sobre toda la oblea para proteger al silicio de la capa de nitruro de silicio que se deposita para la oxidación selectiva. A continuación se define el área activa del dispositivo mediante una máscara de fotorresistencia. Seguidamente se implanta a ambos lados de la zona activa el *channel-stop*. El nitruro no cubierto por la máscara se elimina mediante un proceso de grabado (etching). La oblea se introduce entonces en un horno de oxidación donde sobre las zonas no cubiertas por el nitruro se crece una capa gruesa de óxido (0.5 a 1 μm), que será el responsable del aislamiento eléctrico del dispositivo.

2.- El segundo paso es el crecimiento del óxido de puerta y el ajuste de la tensión umbral. Para ello se elimina toda la capa de nitruro y óxido que cubre la zona activa, y se crece un óxido delgado con un espesor de unos cientos de angstroms. Para ajustar la tensión umbral se implanta el canal con iones de impurezas. En el caso de un transistor en el modo de realce (tensión umbral positiva) se implanta el canal con átomos de boro, hasta que la tensión umbral tenga un valor determinado. Para el caso de transistores canal *n* de deplexión (tensión umbral negativa) se implanta el canal con átomos de arsénico.

3.- El tercer punto es la formación de la puerta. Para ello se deposita una capa de polisilicio y se dopa fuertemente mediante difusión o implantación de fósforo hasta obtener una resistencia laminar típica de 20 o 30 Ω/□. Esta resistencia es adecuada para transistores con longitud de canal mayor de 3 μm. Para canales más cortos se utilizan metales refractarios o siliciuros metálicos para reducir la resistencia laminar por debajo de 1 Ω/□.

4.- El cuarto paso es la formación de la fuente y drenador. El polisilicio que cubre la puerta sirve cómo máscara a la implantación de arsénico para formar la fuente y drenador que estarán autoalineadas con respecto a la puerta. El único solapamiento existente entre puerta y fuente y drenador es debido a la difusión lateral de los iones implantados. Se ha comprobado experimentalmente que usando iones de baja energía es del orden de 50 Å.

5.- El último paso es la metalización. Antes de realizarla, se cubre todo el dispositivo con una capa de óxido dopado con fósforo que por un lado proporciona aislamiento y por otro da a la superficie una topografía suave que evite saltos bruscos y pueda provocar la ruptura de la capa de metalización. El contacto de puerta se realiza normalmente fuera de la zona activa para evitar posibles daños del óxido delgado de puerta en el proceso de fabricación

En la fabricación del transistor NMOS, hay seis procesos de formación de láminas delgadas, cuatro operaciones litográficas, tres implantaciones iónicas, y cuatro operaciones de grabado.

Como ejemplo, la siguiente figura muestra la sección transversal de transistor NMOS una vez finalizada la secuencia de procesos.

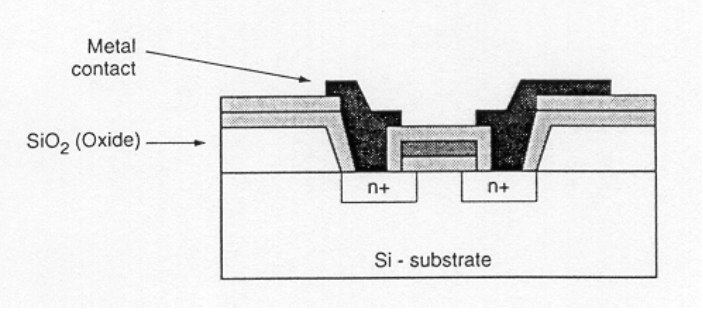


Figura 13.4.1 Sección transversal de un transistor NMOS.

13.5 Tecnología CMOS

En un inversor CMOS ideal no hay disipación de potencia en reposo ya que ninguno de los dispositivos conduce. Éstos lo hacen únicamente cuando se produce la transición entre estados, por lo que únicamente en régimen dinámico es cuando se produce disipación de potencia. Esto hace que la potencia media disipada por el inversor CMOS sea muy pequeña (del orden de nanowatios). Cuando el número de

componentes por chip aumenta, la disipación de potencia se convierte en uno de los principales agentes limitadores (más que el espacio) del número de dispositivos que pueden integrarse. Por esta razón es muy utilizada la tecnología CMOS (aunque un inversor CMOS ocupe más espacio que un inversor NMOS).

Asociado con todos los circuitos CMOS aparece un problema inherente denominado latchup. Este problema está relacionado con la aparición de transistores bipolares parásitos. Puede formarse un transistor npn con la difusión n^+ de fuente o drenador cómo emisor, el pozo *p* como base y el pozo *n* cómo colector. De la misma forma puede observarse la formación de un transistor pnp con la difusión p^+ de fuente o drenador como emisor y con los pozos *n* y *p* cómo base y colector respectivamente. Los dos transistores bipolares parásitos pueden acoplarse y actuar como un tiristor. En condiciones normales de operación, todas las uniones *pn* están polarizadas en inverso. Sin embargo, si por algún motivo los dos transistores bipolares entran en la región activa, el dispositivo posee una gran realimentación positiva (si el producto de las betas es mayor que la unidad) de manera que los transistores conducen produciéndose la ruptura del transistor MOS.

Para que ocurra el latchup, una de las uniones debe estar polarizada en directo fluyendo corriente a través de los transistores. Esta corriente puede proceder de una gran multitud de causas, como por ejemplo la aplicación de una tensión a una de las entradas superior a la de alimentación.

Para evitar el latchup se debe reducir la ganancia de corriente de los dispositivos parásitos. Una forma de conseguirlo es utilizar dopado de oro o irradiación de neutrones para reducir la vida media de los portadores minoritarios. Sin embargo este procedimiento es difícil de controlar. Un procedimiento más efectivo es usar el aislamiento de trinchera (trench isolation). En este caso los transistores bipolares quedan físicamente aislados y por lo tanto se elimina el latchup.

La tecnología CMOS tiene que proporcionar transistores canal *n* y transistores canal *p* sobre la misma oblea. Ahora bien, el substrato de un transistor NMOS es tipo *p*, mientras que el de PMOS es tipo *n*. El principal inconveniente a salvar por la tecnología CMOS es el de proporcionar los dos tipos de substrato. Este problema se resuelve fabricando por implantación o difusión en el substrato de la oblea (que será tipo *n* o tipo *p*) un pozo de conductividad contraria al del substrato. Para ello hay que añadir una concentración de dopante mayor que la existente en el substrato. Si el substrato es tipo *n*, hay que añadir una concentración de aceptadores en la región del pozo $N_A > N_D$ para que el pozo tenga conductividad tipo *p*. Debido a la alta concentración de impurezas que hay que añadir en el pozo, la movilidad del canal quedará degradada ya que ésta depende de la concentración total de dopantes ($N_D + N_A$). Para evitar este problema, se utiliza un substrato muy poco dopado, y se realizan dos pozos, cada uno de ellos con un tipo de conductividad.

En este caso como no es necesaria la compensación de dopantes, la movilidad será mayor.

Dependiendo de la selección inicial (sustrato) el proceso de fabricación CMOS se puede catalogar como procesos de pozo *n*, pozo *p* o pozo gemelo. Este último es más complejo pero el más flexible en la optimización de los dispositivos de canal *n* y *p*. Se ha elegido un proceso CMOS de pozo *n* para mostrarlo aquí, ya que se puede extender fácilmente al caso de tecnología BiCMOS. A lo largo de la explicación se mostrará con figuras la evolución de la estructura hasta llegar al inversor CMOS.

Se necesitan un mínimo de 7 máscaras para completar los dispositivos, no obstante, en la mayor parte de los procesos CMOS se necesitan máscaras adicionales como por ejemplo una segunda capa de polisilicio para la fabricación de capacidades y también en el caso de varios niveles de interconexiones metálicas para conseguir una alta densidad de integración. La inclusión de estos procesos aumentaría el número total de máscaras a más de diez.

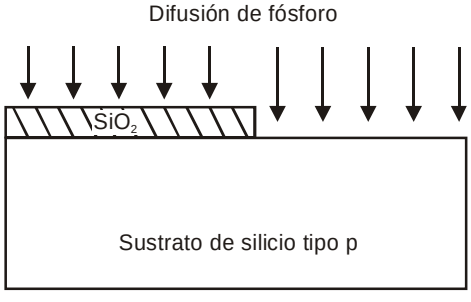


Figura 13.5.1 Máscara 1, difusión de pozo n.

El proceso se inicia con la difusión del pozo *n*. El pozo *n* es necesario siempre que se fabriquen transistores MOS de canal *p*. Se crece una gruesa capa de dióxido de silicio sobre aquellas regiones que se quieren proteger de la difusión de fósforo. Por lo general se emplea fósforo en las difusiones profundas dado que tiene un alto coeficiente de difusión y, en consecuencia, puede difundirse más rápidamente en el sustrato de lo que podría hacerlo el arsénico.

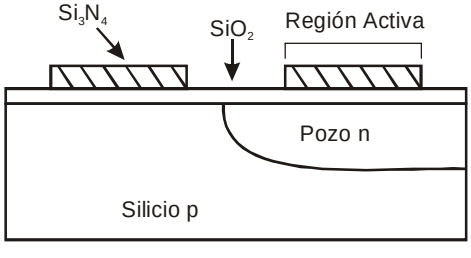


Figura 13.5.2 Utilización de la técnica LOCOS para el crecimiento del óxido. Las zonas cubiertas con Si_3N_4 definen la región activa de los dispositivos.

El segundo paso es definir una región activa (región donde se localizan los transistores) mediante una técnica llamada oxidación local (LOCOS). Se deposita una capa de nitruro de silicio (Si_3N_4) sobre el pozo *n* y otra sobre el pozo *p*. Las regiones cubiertas por el nitruro no se oxidarán de modo que después de un tiempo prolongado de oxidación húmeda aparece un óxido de campo grueso en las regiones situadas entre los transistores. Este óxido grueso es necesario para aislar transistores. También permite que las líneas de interconexión se tracen en la parte superior sin que se formen inadvertidamente canales de conducción en la superficie del silicio.

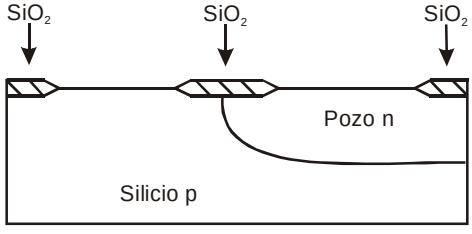


Figura 13.5.3 Resultado de aplicar el proceso de oxidación local LOCOS. Los óxidos gruesos sirven para aislar eléctricamente los dispositivos.

El siguiente paso es la formación de la puerta de polisilicio. Este es uno de los puntos críticos del proceso CMOS. La delgada capa de óxido en la región activa se elimina primero usando un grabado húmedo, seguido por el crecimiento de un óxido muy delgado y de gran calidad en la puerta. De manera rutinaria en los procesos actuales de 0.18 y 0.25 micras se hace uso de grosores de óxido de sólo 100Å. Se deposita una capa de polisilicio generalmente dopado con arsénico (poly tipo n). La fotolitografía es muy exigente en este paso puesto que se requiere una resolución muy fina para conseguir reducir al máximo la longitud de canal del transistor MOS. Esta distancia está representada por el tamaño de la franja más estrecha de polisilicio que se pueda definir.

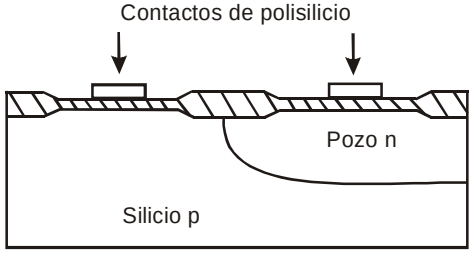


Figura 13.5.4 Formación de las puertas de polisilicio.

La puerta de polisilicio es una estructura que se alinea por si sola y se prefiere sobre al uso de puertas metálicas. Se emplea un implante de arsénico en elevadas dosis para formar las regiones de fuente y drenador n+ de los MOSFETs canal *n*. El contacto de polisilicio también actúa

como barrera para este implante para proteger la región del canal. Se puede usar una capa de material fotorresistente para impedir que las impurezas de As alcancen el transistor de canal p. El óxido de campo de elevado grosor detiene el implante e impide que se formen regiones n+ fuera de las regiones activas.

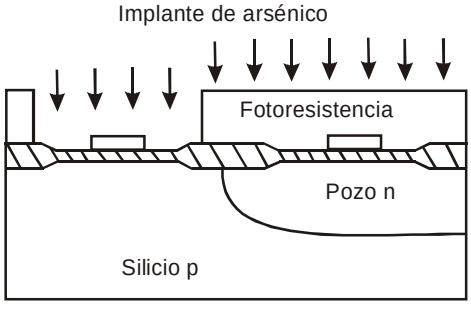


Figura 13.5.5 Implantación de As para formar regiones de fuente y drenador en el transistor MOS de canal *n*.

Se puede realizar un proceso similar de fotolitografía para proteger los MOSFET *n* durante el implante de boro cuando se definen los contactos de fuente y drenador en los MOSFET canal *p*. Nótese que en ambos casos la separación entre las difusiones de fuente y drenador definida como longitud de canal, viene dada sólo por la mascara de puerta de polisilicio, de ahí la propiedad de autoalineamiento.

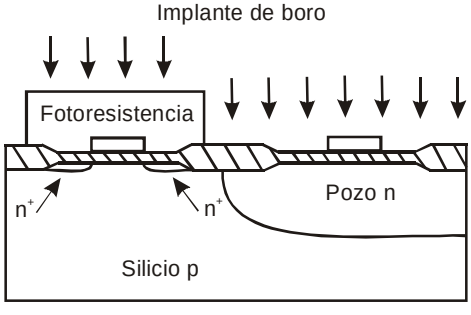


Figura 13.5.6 Implantación de boro para formar regiones de fuente y drenador en el transistor MOS de canal *p*.

Antes de que se abran los huecos para realizar los contactos, se deposita en toda la estructura una gruesa capa de óxido mediante un proceso denominado Chemical Vapor Deposition (CVD). Se emplea una fotomáscara para definir la abertura de ventana de los contactos seguida por un grabado de óxido húmedo o en seco.

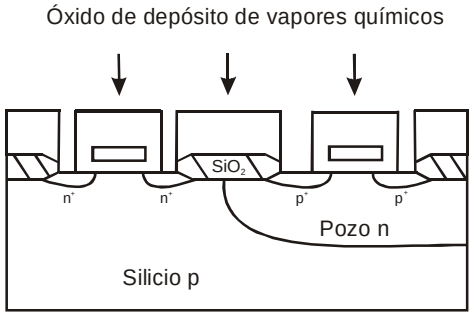


Figura 13.5.7 Deposición de una gruesa capa de óxido mediante CVD. Abertura de las ventanas donde se realizan los contactos a los diferentes terminales de los transistores.

A continuación se vaporiza o metaliza por bombardeo iónico una delgada capa de aluminio sobre la oblea. Se emplea un paso final de enmascaramiento y grabado para formar la interconexión. El paso final antes del empaquetamiento y conexión es la pasivación de la superficie mediante un tratamiento con soluciones ácidas para eliminar partículas y residuos. Por lo general se deposita una gruesa capa de óxido mediante CVD o cristal pirex sobre la oblea que actúa como protección.

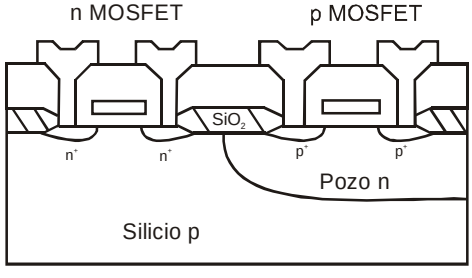


Figura 13.5.8 Metalización para formar los contactos de fuente/drenador en los transistores *p* y *n*.

13.6 Tecnología SOI

El silicio sobre aislante (SOI) es un nuevo material para la fabricación de chips que reemplaza a las obleas tradicionales de silicio. Este substrato está compuesto de tres capas diferenciadas. Primero una delgada lámina de silicio que puede ir desde unos pocos nanometros hasta varias micras. A continuación una capa de aislante, normalmente óxido de silicio con un grosor variable y finalmente la lámina más gruesa de silicio que actúa como material que da soporte mecánico a toda la estructura. Los dispositivos se fabrican en la capa superficial de silicio. Cada transistor se encuentra aislado del resto gracias al óxido enterrado. Esta característica permite evitar el fenómeno de latchup y fabricar más transistores por cm^2. Además se ha demostrado una importante mejora en sus prestaciones ya que pueden trabajar a menores tensiones, aumentar la velocidad de conmutación y son menos vulnerables al efecto de las partículas cósmicas

y efectos de canal corto (SCE). Todas estas mejoras se consiguen sin necesidad de alterar los procesos que tradicionalmente se han seguido en tecnología CMOS. El cambio fundamental se produce en el punto de partida ya que las obleas son completamente diferentes. Se han desarrollado muchas técnicas para conseguir una lámina muy delgada de silicio sobre aislante con buenas propiedades (espesor uniforme en toda la superficie, baja densidad de defectos en el volumen y en las interfases, etc). En sus inicios se realizaba un crecimiento epitaxial de silicio sobre una oblea cubierta de aislante (técnicas homoepitaxial y heteroepitaxial). Otras técnicas se basaban en la cristalización de una lámina delgada de silicio previamente fundida (recristalización láser o por haces de electrones). Posteriormente se utilizó una implantación de oxigeno sobre silicio para crear el óxido enterrado (SIMOX). No obstante, estos procedimientos no proporcionaban regiones activas de calidad comparable a las obleas de silicio puro y además los costos seguían siendo elevados ya que la producción y la demanda seguían siendo bajos. Hoy en día se utilizan técnicas de *wafer bonding*, pegado de obleas, que proporcionan muy buenos resultados ya que mejoran la calidad al tiempo que reducen costes. Son precisamente estos procesos de fabricación de obleas los que van a ser comentados a continuación. Partiendo del substrato de SOI la secuencia de fabricación de transistores es similar a lo que hemos comentado para tecnología CMOS.

ELTRAN, Epitaxial Layer Transfer

El silicio poroso se forma mediante una reacción electroquímica cuando el silicio constituye el ánodo de una celda electrolítica con ácido fluorhídrico (HF) como electrolito. Esta técnica utiliza el hecho de que el silicio poroso es mecánicamente débil pero mantiene la estructura cristalina del substrato en el que se formó. Mediante un recocido a elevadas temperaturas en ambiente de hidrógeno se sellan los poros en la superficie de la oblea. Sobre este silicio poroso sellado se crece epitaxialmente una lámina de silicio y a continuación un óxido térmico. En este punto la oblea se une con otra que actuará como soporte mecánico. Puesto que el silicio poroso es mecánicamente frágil puede romperse con facilidad con, por ejemplo, un chorro de agua a presión. Una mejora posterior ha sido el uso de dos láminas de silicio poroso con diferente morfología. Puesto que se produce una tensión muy fuerte en la interfase entre las dos láminas, el chorro de agua produce un corte limpio entre estas dos interfases disminuyendo la rugosidad en la superficie. El silicio poroso que permanece en la superficie de la oblea se elimina y nos encontramos con la superficie del silicio sobre aislante (SOI) que nuevamente se somete a un proceso de recocido a 1100ºC en un ambiente rico en hidrogeno. La oblea sobrante se puede reutilizar nuevamente con el consiguiente ahorro de costes. La tecnología ELTRAN se ha empleado con éxito en obleas de 300mm (12 in) y se han conseguido espesores de silicio inferiores a 30nm con una calidad comparable a otras más gruesas.

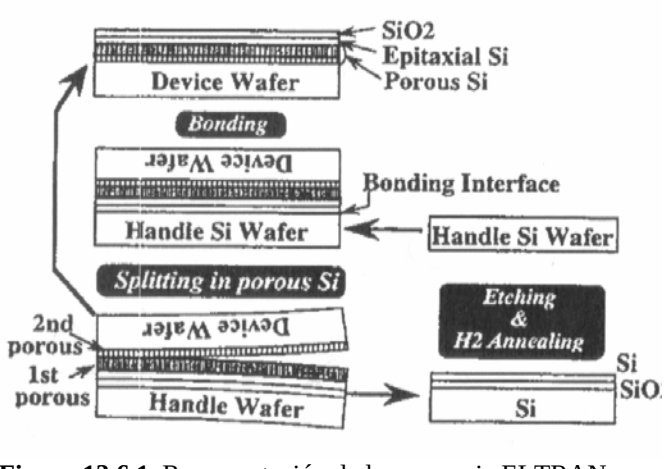


Figura 13.6.1 Representación de la secuencia ELTRAN para la fabricación de obleas SOI.

Smart Cut

Al implantar iones de hidrógeno en una dosis superior a $5\times10^{16}cm^{-2}$ se producen microcavidades en la red de silicio. Si esa oblea implantada se calienta a temperaturas entre 400 y 500 ºC los átomos de hidrógeno pasan a formar moléculas H_2 haciendo que la presión aumente hasta un punto de fractura. Para una dosis implantada superior a $10^{17}cm^{-2}$ se forman pequeñas burbujas en la zona implantada incluso sin el tratamiento térmico. La versión comercial del proceso que se acaba de comentar se denomina Smart Cut y se desarrolló en el CEA-Leti de Grenoble (Francia). Esta tecnología la comercializa la empresa SOITEC y la primera familia de obleas se denominó UNIBOND.

La secuencia del proceso necesario para conseguir obleas SOI mediante el proceso Smart Cut es la siguiente. La oblea inicial se oxida con el espesor deseado. Este óxido se convertirá posteriormente en el óxido enterrado de los dispositivos resultantes. El siguiente paso es la implantación de hidrógeno a través del óxido en una dosis superior a $5\times10^{16}cm^{-2}$. Después de la implantación la oblea soporte y la oblea semilla se limpian cuidadosamente para eliminar cualquier partícula o contaminante y hacer las dos superficies hidrofílicas. Las dos obleas se alinean y funden para formar un único cuerpo. A continuación estas obleas se introducen en un horno calentado en un rango de temperaturas de entre 400 y 600ºC que produce la separación entre obleas a lo largo del implante de hidrógeno que es la zona más frágil mecánicamente. La superficie de las obleas resultantes presenta una rugosidad de unos pocos nanometros. Un proceso posterior de pulido consigue la misma rugosidad superficial que una oblea de silicio convencional. La oblea semilla puede reutilizarse de nuevo reduciendo el costo final de la oblea SOI. Esta oblea semilla es la que proporciona la lámina de silicio y por este motivo debe ser de gran calidad mientras que la segunda oblea actúa únicamente como

soporte mecánico por lo que no se necesita gran calidad. El hecho de definir el espesor de lámina de silicio mediante la energía de implantación permite un control mucho más preciso del que es posible conseguir con cualquier proceso mecánico o químico. El espesor del óxido y/o de la lámina enterrada pueden ajustarse en el proceso Smart Cut eligiendo la energía de implante y el tiempo de oxidación. El espesor de la lámina de silicio se mueve en el rango de 5nm a 1.5μm y el espesor del óxido puede ser tan delgado como 2nm. Actualmente se utilizan nuevas técnicas para mejorar los resultados, por ejemplo el uso combinado de hidrógeno y helio se ha demostrado más eficaz en la separación de las obleas.

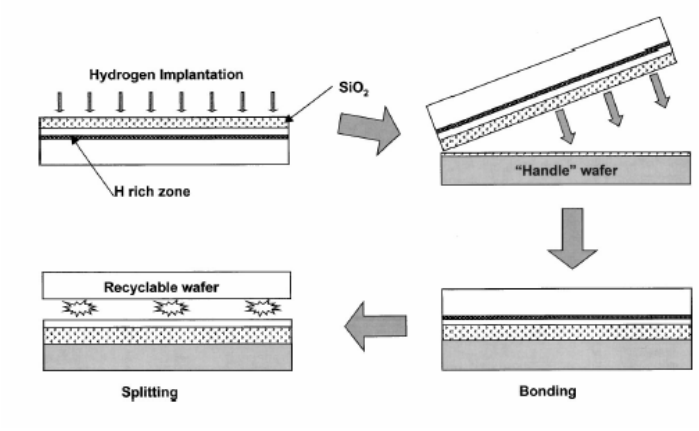


Figura 13.6.2 Secuencia de pasos seguida para fabricar obleas SOI siguiendo la técnica SmartCut.

13.7 Tecnología BiCMOS

Para conseguir una elevada tensión de ruptura en la unión base colector de un transistor bipolar es necesario utilizar una lámina epitaxial muy gruesa (17μm de material con 5Ω-cm para 36V). Si se permiten tensiones de ruptura mucho más bajas (por ejemplo 7V si se trabaja con tensiones de alimentación de 5V) entonces se puede utilizar un dopado elevado en el colector (del orden de 0.5Ω-cm). En este caso es posible aislar lateralmente los diferentes dispositivos fabricados siguiendo la tecnología bipolar mediante capas de óxido gracias a la técnica LOCOS. Esto tiene la gran ventaja de reducir enormemente la capacidad parásita existente entre colector y substrato porque las regiones muy dopadas próximas a la superficie ahora se sustituyen por las capacidades mucho menores de los óxidos de aislamiento. Además, los dispositivos se pueden empaquetar con mayor densidad dentro del chip y la tecnología de fabricación bipolar y CMOS comienzan a parecerse mucho.

La utilización conjunta de transistores bipolares y CMOS es atractiva desde el punto de vista del diseño digital puesto que la alta capacidad de conducción de corriente del transistor bipolar facilita el

manejo de grandes capacidades de carga con tiempos de carga y descarga pequeños. Desde el punto de vista del diseño analógico también es interesante el poder disponer en un mismo circuito de transistores de los dos tipos (altas impedancias de entrada de los transistores MOS y alta ganancia de corriente de los transistores bipolares).

El proceso comienza con el enmascaramiento y la implantación de iones de antimonio para la formación de capas enterradas n^+ dentro del substrato tipo *p*, en las zonas donde se vaya a realizar el transistor bipolar npn o el transistor PMOS. Una segunda implantación de boro se realiza para formar el pozo donde se fabricará el dispositivo NMOS. Se crea una capa epitaxial de 1 µm de espesor para formar los colectores de los transistores bipolares y el canal del transistor PMOS. A continuación se crecen las capas de óxido de campo que sirven para aislar a los diferentes dispositivos. Finalmente se realiza una serie de implantaciones para formar las regiones de base y emisor en los transistores bipolares y de drenador y fuente en los transistores MOS. Durante esta secuencia se crece el óxido de puerta, se crecen las puertas de polisilicio y se ajusta la tensión umbral de los transistores mediante implantación. Posteriormente se depositan las diferentes capas de metalización. El resultado puede observarse en la siguiente figura,

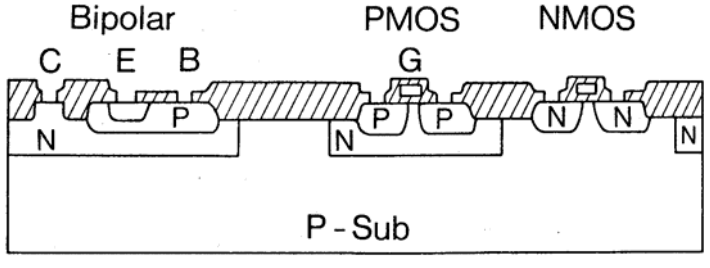


Figura 13.7.1 Deposición de una gruesa capa de óxido mediante CVD. Abertura de las ventanas donde se realizan los contactos a los diferentes terminales de los transistores.

13.8 Tecnología MESFET

Avances recientes en el procesamiento tecnológico del arseniuro de galio han hecho posible una tecnología de circuitos integrados en arseniuro de galio similar a la del silicio.

El arseniuro de galio tiene tres ventajas fundamentales sobre el silicio:

1.- Mayor movilidad de los electrones, lo que se traduce en una menor resistencia serie para una geometría dada, y mayores niveles de corriente.

2.- Mayor velocidad de deriva para un valor aplicado de campo eléctrico, lo que mejora la velocidad de respuesta del dispositivo.

3.- Pueden crecerse capas de arseniuro de galio semiaislante, lo que proporciona la posibilidad de substratos cristalinos aislantes.

Sin embargo, el arseniuro de galio posee también tres inconvenientes importantes frente al silicio:

1.- Vida media de minoritarios muy corta.

2.- Ausencia de un óxido estable y de buena calidad.

3.- Los defectos cristalinos en el arseniuro de galio son muchos órdenes de magnitud mayores que en el caso del silicio.

La escasa vida media de los minoritarios y la ausencia de un óxido estable y de buena calidad ha hecho imposible la fabricación de transistores MOS en arseniuro de galio.

Por lo tanto, la tecnología del arseniuro de galio se ha basado principalmente en los transistores MESFET en la que los portadores mayoritarios son transportados a través de contactos metal semiconductor.

El material inicial para la construcción de un circuito integrado en arseniuro de galio es un substrato semi-aislante de GaAs. Sobre él se crece una lámina *buffer* de GaAs muy puro antes de la deposición de lo que será la lámina activa en si misma. De esta forma se consigue una importante mejora en la movilidad de los portadores al reducir la densidad de defectos en la interfase de la región activa. Encima de la lámina activa se crece otra lámina de GaAs muy dopado con objeto de reducir el valor de la resistencia de acceso. Esta lámina se elimina de la región que queda bajo la puerta mediante un ataque selectivo. Los métodos más frecuentes para conseguir estas láminas de material son la epitaxia en fase de vapor y la epitaxia por haces moleculares.

La fabricación de MESFETs requiere de cuatro pasos básicos: aislamiento, contactos óhmicos, contactos Schottky y finalmente pasivación y metalización. Cada uno de estos pasos involucra áreas localizadas de la pieza semiconductora que se definen mediante fotolitografía convencional o litografía por haces de electrones. Las técnicas de fotolitografía o MBE se explican con detalle en capítulos previos.

El propósito esencial del aislamiento es restringir el área eléctricamente activa de toda la lámina de GaAs a las regiones activas de cada uno de los dispositivos fabricados. La técnica más empleada para los dispositivos discretos se denomina aislamiento en meseta. En esta técnica la región activa del dispositivo se protege con un material fotorresistente y se ataca el resto del material que no ha sido previamente protegido. Normalmente para este propósito se utiliza un ataque químico. El inconveniente principal de esta técnica es que el relieve de la estructura resultante produce graves dificultades para su posterior recubrimiento con contactos metálicos, máscaras u otros materiales. Otro método de conseguir el aislamiento es mediante una implantación iónica. En un caso se implantan iones en las regiones no activas del dispositivo sin un posterior recocido. Puesto que la implantación produce graves daños en la red, la región implantada se convierte en aislante. En el caso de

implantación selectiva, que se utiliza a menudo en la fabricación de circuitos integrados, el área activa se forma implantando iones en la región activa de cada dispositivo y posteriormente desarrollando un recocido para su activación. La principal ventaja de estas dos técnicas es que conserva la planaridad de la estructura resultante.

- Contactos óhmicos

Los contactos óhmicos de fuente y drenador de un MESFET de GaAs deberían tener una característica I-V lineal con una resistencia de contacto mínima. El método más usado para conseguir un buen contacto óhmico es depositar una lámina de AuGeNi por evaporación y entonces formar una aleación con el GaAs. De hecho la formación de contactos óhmicos sobre GaAs es un proceso muy complejo que aún hoy no se comprende completamente. Para contactos óhmicos planos en MESFETs la transición entre el material en volumen y el contacto se caracteriza mediante una resistencia de contacto que normalmente es menor que 0.1 Ωmm.

- Barreras Schottky

Debido a la reducción de dimensiones, la formación de la barrera Schottky es el paso crítico en la fabricación de MESFETs. El método más utilizado en la fabricación de barreras Schottky es depositar aluminio o una aleación de AuPtAu sobre la zona deseada. Como resultado se obtiene una característica I-V que, en condiciones de polarización directa, responde a la siguiente expresión

$$I = A^* T^2 S \exp\left(\frac{qV_b}{kT}\right)\exp\left(\frac{qV_{GS}}{nkT}\right), \qquad (13.6)$$

donde A^* es la constante de Richardson, S el área, V_b la altura de la barrera, n el factor de idealidad y V_{GS} la tensión de puerta. Normalmente V_b es aproximadamente 0.8 y n 1.1. La puerta también se caracteriza por la resistencia de metalización que es un factor que puede limitar el comportamiento del dispositivo.

- Formación de dieléctricos: Interconexiones

Los pasos finales en la fabricación de un MESFET son la formación del dieléctrico y la interconexión. El propósito de la lámina dieléctrica es proteger la superficie de posibles ataques químicos o mecánicos. Estas láminas generalmente se forman mediante una deposición asistida por plasma o *sputtering*. El dieléctrico más utilizado sobre GaAs es el nitruro de silicio (Si_3N_4). En el caso de circuitos integrados, las láminas dieléctricas también se utilizan como aislante entre dos niveles de metalización.

El problema de interconexión es importante en MESFETs de potencia y en circuitos integrados cuando se necesita aumentar la anchura total de la puerta. Son dos las técnicas más empleadas para realizar las interconexiones. La primera se conoce como puentes sobre aire que se fabrican depositando el metal sobre una lámina gruesa del semiconductor que posteriormente se elimina. El segundo método consiste en un ataque

químico a una lámina de GaAs previamente adelgazada. De esta forma se consigue una importante reducción de las capacidades e inductancias parásitas que mejoran el comportamiento del dispositivo.

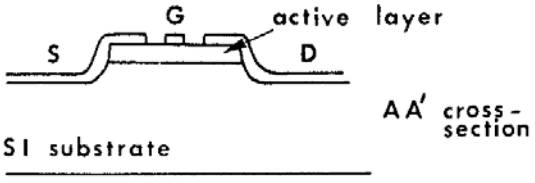


Figura 13.8.1 Imagen de un transistor MESFET que utiliza aislamiento de tipo meseta.

RESUMEN

En este capítulo se ha estudiado la tecnología de fabricación de circuitos integrados. Se ha tratado de explicar brevemente la secuencia de procesos necesarios para fabricar un dispositivo en cada una de las tecnologías seleccionadas basándonos en las explicaciones de capítulos previos dedicados al estudio de la oxidación, difusión, implantación iónica, litografía, etc.

Se ha hecho un recorrido por las diferentes tecnologías que históricamente han jugado un papel importante. Algunas como la bipolar y NMOS se encuentran en un claro declive ya que sus años de apogeo pasaron y han sido sustituidas por la CMOS que actualmente domina el mercado con claridad. Para algunas aplicaciones específicas podemos encontrar la BiCMOS. Muy importante en nuestros días es el silicio sobre aislante (SOI) que está conquistando rápidamente una importante cuota de mercado debido a las ventajas que le confiere el aislante enterrado. Existe la posibilidad de fabricar dispositivos con materiales diferentes al silicio como por ejemplo el GaAs, AlGaAs, InP, etc. Sin embargo todos ellos ocupan nichos de mercado muy restringidos a aplicaciones específicas (comunicaciones aerospaciales o militares fundamentalmente). Nunca han tenido éxito debido a sus elevados costes y a la constante mejora de la tecnología de silicio que sigue inexorable la Ley de Moore.

Por supuesto existen variantes a los procedimientos aquí mostrados ya que por ejemplo no sería lo mismo fabricar un dispositivo de potencia que otro diseñado para actuar a frecuencias elevadas. No obstante, conocidos los procesos básicos la alteración de la secuencia no implica una mayor dificultad conceptual.

REFERENCIAS

[1] S. Wolf. *Silicon Processing for the VLSI Era, Volume 4.* Lattice Press, 2002.

[2] Gray, Hurst, Lewis and Meyer. *Analysis and Design of Analog Integrated Circuits*. John Wiley and Sons, 4th Edition, 2001.

[3] A.S. Sedra and K.C. Smith. *Microelectronic Circuits*. Oxford University Press, New York, 1998, p. 139.

[4] G. K. Celler, S. Cristoloveanu, "Frontiers of silicon-on insulator"; J. Appl. Phys., vol. 93, no.9, pp. 4955-4978, 2003.

[5] http://www.soitec.com

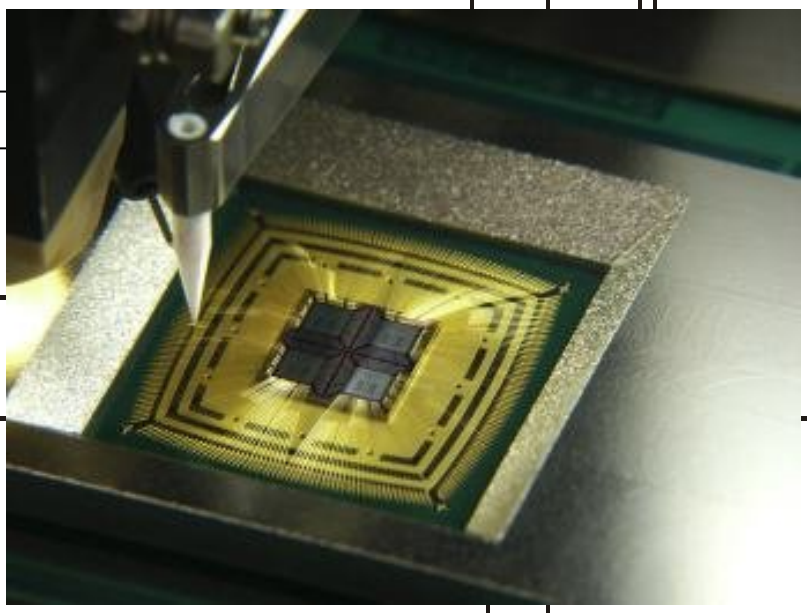
[6] http://www.sigen.com (Technology; Technical References)

[7] http://www.canon.com/technology/detail/device/soi

14

Capítulo

Aspectos industriales de la fabricación de



Proceso de cableado del chip

Índice

Objetivos

- Describir las diferentes técnicas para el análisis y resolución de las problemas de la tecnología ULSI.
- Estudiar el proceso de encapsulado de un chip.
- Ver los fenómenos que afectan al rendimiento de fabricación y a la fiabilidad.

Palabras clave

Nomarski
SEM
TEM
Auger
NAA
RBS
SIMS
XRF
XPS
Reflactancia laser
Channeling
X ray diffraction
TED
EBIC
Empaquetado cerámico
Empaquetado plástico
Back grinding
Dicing
Chip bonding
Pegado eutectico
Wire bonding
Empaquetamiento TCP
Empaquetamiento Flip-Chip
Rendimiento
Fiabilidad

14.1 Técnicas de diagnóstico

En este apartado veremos los diferentes métodos instrumentales para el análisis y resolución de los problemas de la tecnología VLSI.

Existen diferentes métodos instrumentales para la detección de problemas en la fabricación VLSI. Estos métodos se usan fundamentalmente para cuatro tipos de análisis:

1) La determinación de la morfología
2) Análisis químico
3) Determinación de la estructura cristalográfica y de las propiedades mecánicas
4) Mapeo eléctrico de los lugares con roturas y fugas

Método instrumental	Acrónimo	Determinación de la morfología	Análisis Químico	Estructura cristalográfica y propiedades mecánicas	Mapeo Eléctrico
Auger electron Spectroscopy	AES		X		
Electron Bean induced current microscopy	EBIC				X
Laser reflectance	LR			X	
Neutron activation Analysis	NAA		X		
Normarski interference contrast optical microscopy		X			
Rutherford backscattering spectroscopy	RBS		X	(X)	
Scanning electron microscopy	SEM	X	(X)		(X)
Secondary ion mass spectroscopy	SIMS		X		
Transmission electron diffraction	TED			X	
Transmission electron microscopy	TEM	X	(X)	X	(X)
Voltage contrast microscopy	VC				X
X-ray diffraction	XRD			X	
X-ray emission microscopy	XES		X		
X-ray fluorescence	XRF		X		
X-ray photoelectron spectroscopy	XPS, ESCA		X		

Tabla 14.1.1: Aplicación de los métodos instrumentales para análisis de los problemas de la VLSI

En la Tabla 14.1.1 se muestra los diferentes métodos instrumentales existentes y para que diagnostico se utiliza cada uno. Las cruces indican que tipo de análisis se puede realizar con cada método instrumental, las cruces entre paréntesis indican que se puede utilizar esa técnica experimental pero con equipo accesorio especial.

En algunos procedimientos analíticos descritos en la las muestras son bombardeadas con un haz de electrones o rayos X y el análisis requiere la medida de la radiación resultante.

En la Tabla 14.1.2 se muestran los procesos basados en estas iteraciones y los rangos de energía típica de las radiaciones incidentes y secundarias:

Rayo incidente		**Radiación secundaria**				
			Electrón		X-ray	
Radiación	Energía E_0 (keV)	Tipo	Energía (eV)	Procedimiento analítico	Energía	Procedimiento analítico
Electrón	2-10	Auger	20-200	AES		
	2-40	Secundaria	<10	SEM(VC)		
	2-40	Back-scattared	$<E_0$	SEM(BS)		
	20-200				$<E_0$	XES
X-ray	<2	Primary ionized	20-200	XPS		
	<50				$<E_0$	XFR

Tabla 14.1.2: Tipos de radiación resultante del bombardeo de la superficie de una muestra y métodos instrumentales basados en esas interacciones

14.1.1 Determinación morfológica

Con este análisis se examina las formas de partes relevantes como los bordes de las líneas grabadas, proximidad entre estas, alineamiento. Estas características se pueden examinar por microscopia óptica, SEM (scaning electrón microscopy) y TEM (transmisión electrón microscopy). Los aumentos máximos de cada técnica son aproximadamente de 1000X, 50000X y 500000X respectivamente. Puesto que los rangos de magnificación se superponen pocas características morfológicas podrán escapar del escrutinio.

14.1.1.1 Microscopia optica de interferencia de contraste de Nomarski

La microscopia óptica de interferencia de contraste de Nomarski es la forma mas usual de microscopia óptica para resolver problemas de proceso en la VLSI. Con este método la forma de la superficie con diferentes alturas aparece en diferentes colores o diferentes niveles de grises. Este contraste se consigue separando un rayo de luz en dos desplazados una pequeña distancia en la superficie de la muestra seguido

de una reflexión y preconstitución de los rayos reflejados. La longitud del camino óptico cambia debido a la presencia de escalones y cambios en el índice de refracción (debido a una frontera de fase). Estos cambios en la longitud de los caminos ópticos producen un cambio de contraste en el haz reconstituido, lo cual aparece en la imagen del microscopio.

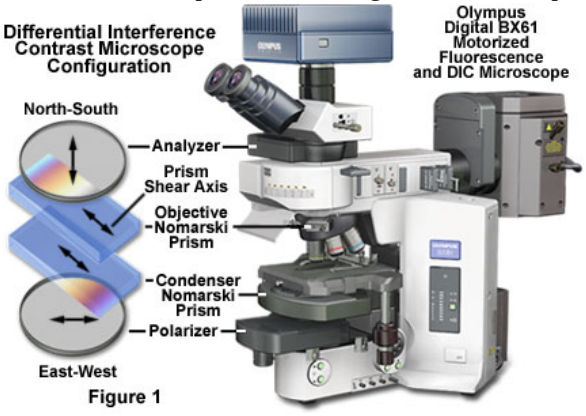


Figura 14.1.1: Configuración de un microscopio óptico de interferencia de contraste.
http://www.microscopy.fsu.edu/primer/techniques/dic/dicintro.html
http://www.olympus.co.jp/en/insg/ind-micro/product/semicon.cfm

Figura 14.1.2: Microscopio Optico Leitz Wetzlar con interferencia de contraste de Nomarski
http://www.staff.ncl.ac.uk/k.vassilevski/EECE-CR/Leitz.htm

La resolución de un microscopio óptico viene dada por la longitud de onda de la fuente de luz λ y la apertura numérica (NA) de la lente del objetivo:

$$r = \frac{0.61\lambda}{NA} \tag{14.1}$$

El limite de la resolución de un microscopio óptico es aproximadamente de 0.25μm. Si a resolución visual con el ojo humano se toma 0.1mm, la perdida de la claridad de imagen puede empezar a

detectarse por encima de los 400 aumentos y el limite superior para una ampliación útil esta comprendido entre los 1000 y 2000 aumentos. Para muchos de los problemas que requieren el análisis de la morfología de la superficie una resolución lateral de 0.25 o incluso 1μm es bastante aceptable. Un nivel de resolución vertical así es también útil , permite la clara identificación de zonas con grosores hasta de 200Å.

El sistema de microscopia por contraste de interferencia de Nomarski se muestra en la Figura 14.1.3:

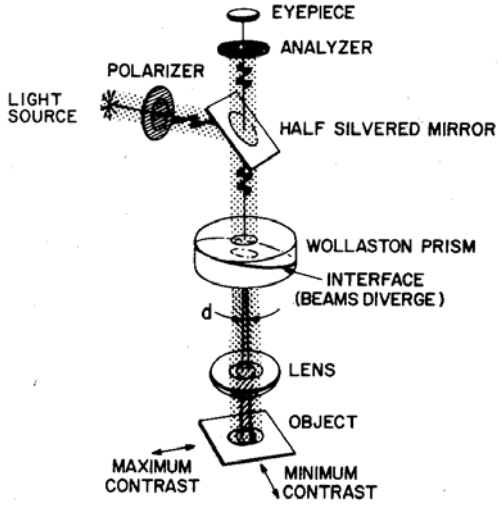


Figura 14.1.3: Esquema de un microscopio optico de contraste de interferencia de Nomarski “libro VLSI Technology de Sze”

La luz pasa a través de un polarizador y es reflejada para abajo, hacia unos cristales birrefrigentes que juntos forman un prisma Wollaston, esto es, un prisma en el cual la luz se separa en dos componentes polarizadas perpendiculares que se mueven a diferentes velocidades con una divergencia angular d. Después de emerger del prisma y reflejarse en la muestra, los dos rayos se recombinan al pasar nuevamente por el prisma en dirección opuesta. El haz reconstituido pasa entonces a través de un analizador donde sus cambios de intensidad son observados.

14.1.1.2 SEM Scaning Electrón Microscopy

La técnica de microscopia por escáner de electrones SEM es un método analítico estándar en los laboratorios VLSI, principalmente porque proporciona una mayor resolución espacial y mayor profundidad de campo comparada con la microscopia óptica, además de proporcionar información química del espectro de rayos x generado por el bombardeo de electrones. Se pueden conseguir resoluciones mejores de 100Å en condiciones optimas y profundidades de campo de 2 a 4μm a diez mil aumentos y de 0.2 a 0.4mm para cien aumentos.

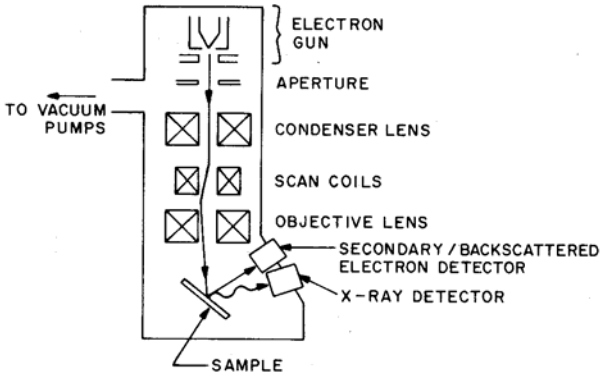


Figura 14.1.4: Esquema de un microscopio electrónico de barrido "libro VLSI Technology de Sze"

La Figura 14.1.4 muestra el dibujo de un esquema de un microscopio SEM. Este esta formado por un cañón de electrones, que consiste en un filamento de tugsteno o LaB_6 que genera los electrones, estos son acelerados a energías de 2 a 40 keV. Una combinación de lentes magnéticas y scan coils proporciona un haz de diámetro pequeño es cual es rastreado a lo largo de la superficie de la muestra. El bombardeo de electrones produce tres tipos útiles de radiación : rayos X, electrones secundarios y electrones dispersados hacia tras (backscattered electrons).

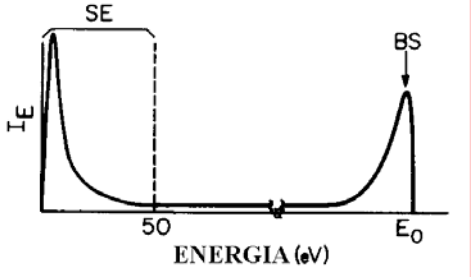


Figura 14.1.5: Espectro de energias de los electrones emitidos "libro VLSI Technology de Sze"

La Figura 14.1.5 muestra el espectro de energías de los electrones emitidos por la muestra que ha sido bombardeada con un haz de electrones. Una gran parte del espectro lo forma electrones con energías inferiores a 50 eV teniendo un máximo para menos de 5eV. A esta emisión de electrones se le denomina **emisión secundaria**. El otro pico de emisión de electrones se encuentra para energías cercanas a la energía de los electrones incidentes E_0 y esta emisión se denomina **electrones dispersados hacia atras** (backscattered electrons).

La corriente de los electrones secundarios y dispersados hacia tras se utiliza para modular la intensidad de un haz de electrones en un tubo de rayos catódicos (CRT). Ya que el haz de electrones del CRT se mueve sincronizado con el haz de electrones de rastreo del SEM , el haz del CRT produce una imagen de la superficie de la muestra cuyo contraste

esta determinado por las variaciones del flujo de electrones secundarios y dispersados hacia tras. La emisión de rayos X es útil para el análisis químico.

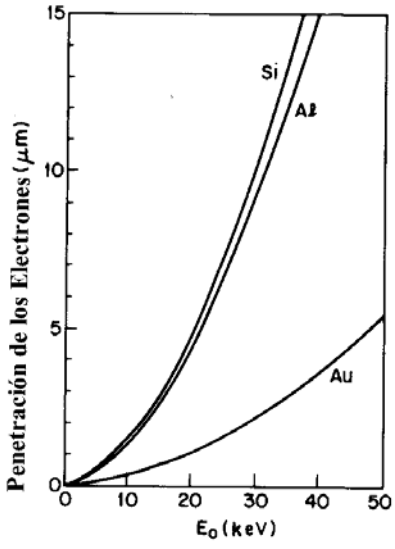


Figura 14.1.6: Penetración de los electrones en función de la energía de los electrones incidentes para diferentes materiales

El haz de electrones incidentes experimenta múltiples colisiones al penetrar en la muestra. Los electrones que no son reflejados finalmente se detienen después de atravesar una distancia R que puede ser calculado o medida. La Figura 14.1.6 muestra el cálculo de las distancias que recorren los electrones para el silicio el aluminio y el oro. Estas distancias se incrementan cuanto menor es el número atómico de la muestra y mayor es la energía E0 del haz de electrones incidentes. La trayectoria de los electrones varía con cada colisión lo que causa que el estrecho rayo incidente se disperse a medida que penetra en la muestra.

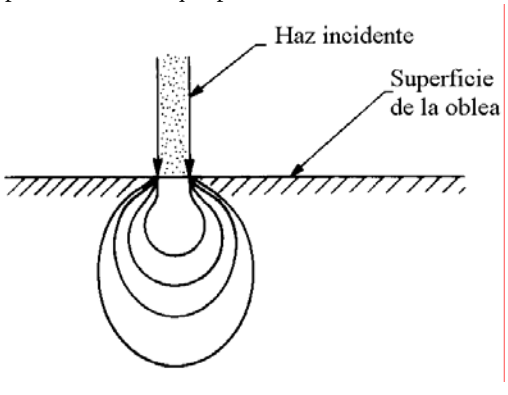


Figura 14.1.7: Penetración máxima de los electrones incidentes para cuatro energías deferentes

La Figura 14.1.7 muestra la penetración máxima de los electrones incidente para cuatro haces de energías diferentes. La penetración viene dada por una serie de círculos los cuales incrementan su profundidad y anchura a medida que la energía aumenta.

Para la incidencia normal sobre una muestra el área que contribuye en la emisión de electrones dispersados hacia tras, es un disco con un diámetro aproximadamente equivalente a la profundidad de los electrones. La resolución de la imagen de electrones dispersados hacia tras es mejor cuanto menor es el espesor de la muestra por ejemplo para capas de metal o óxidos delgados. La resolución de las imágenes formadas por la emisión secundaria de electrones viene determinada por el tamaño lateral del material dentro del cual la emisión secundaria es generada para una profundidad inferior que la profundidad de escape. La profundidad de escape de los electrones en un metal alcanza un mínimo de 4Å a 70eV y se incrementa cuando disminuye la energía 25Å para 10eV. La profundidad de escape es mayor de 50Å en los aislantes. La resolución lateral de la emisión secundaria para una superficie plana viene dada por tanto por el diámetro del haz incidente mas un incremento lateral debido al camino libre principal de los electrones.

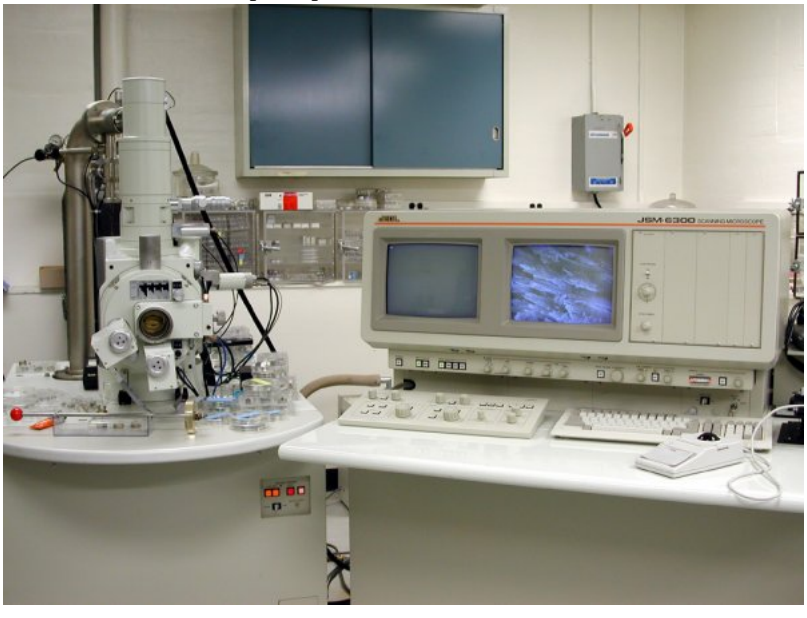


Figura 14.1.8: Microscopio Electrónico de Barrido JEOL 6300V

http://www.med.sc.edu:89/IRF/EQUIP.HTM
http://www.jeol.com/wi/wi.html

El contraste de ambas emisiones electrones dispersados hacia tras y secundaria depende de las variaciones de flujo de electrones que llegan al detector. El rendimiento la emisión de los electrones dispersados hacia tras depende del numero atómico Z. Así que se produce un contraste entre materiales con distintos números atómicos. Este es de un

6,7% entre el aluminio y el silicio. Por tanto la detección de electrones dispersados hacia tras es útil para la detección de partículas de aluminio sobre fondo de silicio.

La emisión secundaria de electrones no depende tanto del numero atómico Z. En este caso el rendimiento depende de la diferencia de función trabajo de los materiales. Con esta emisión se puede claramente diferenciar las regiones de oxido, metales y semiconductores. Una segunda fuente de contraste en la emisión secundaria se debe a la dependencia del rendimiento con una curvatura de la superficie. Las superficies cuya pendiente difieren pueden se claramente distinguidas.

La resolución espacial por tanto depende de la superficie de la muestra que contribuye a las emisiones secundaria y dispersada hacia tras, de los cambios locales de fase, composición y orientación de la muestra. Los cuales como se ha explicado influyen en el flujo de electrones de las dos emisiones.

El análisis de los circuitos VLSI mediante SEM tiene tres problemas: la contaminación de la muestra, el daño que produce el haz de electrones y las cargas que se inducen en la superficie durante el análisis. La mayor parte de contaminación se produce por polimerización de hidrocarbono producida cuando el haz de electrones choca con la superficie. Aunque los microscopios modernos tienen unas buenas bombas para mantener un vacío de unos 10^{-6} Torr esta contaminación no puede ser evitada. El segundo problema es el daño que causa los electrones en lo óxidos. La radiación de electrones produce cargas positivas en el oxido y trampas en la interfaces las cuales se pueden evitar si se mantiene una energía baja para que no penetren los electrones en las zonas activas como los oxido de puerta. Una vez que se han formado estos defectos pueden ser eliminados por recocido a temperaturas entre 400 y 550º C

El tercer problema es la carga inducida en superficies aislantes. Esto sucede cuando las energías del haz incidente están por encima del punto de cruce de rendimiento de la emisión secundaria. La superficie se carga negativamente enturbiando la trayectoria del haz de electrones incidente y degradando la imagen. Una forma de evitar esto es usar un haz incidente de baja energía.

14.1.1.3 Microscopia electrónica de transmisión (Electrón Transmision Microscopy TEM)

La microscopia por transmisión de electrones (TEM) es una herramienta habitual para resolver problemas de la tecnología VLSI que requieren una alta resolución espacial. TEM ofrece una resolución de 2Å. En un microscopio de transmisión de electrones un haz de electrones atraviesa la muestra de fina capa y forma una imagen que proporciona las características morfológicas y cristalográficas de los componentes de la misma. Los microscopios TEM comerciales utilizan haces de electrones con energías entre 60 y 350 KeV. Cuanto mayor es la energía mayor será

el espesor de la muestra que se podrá analizar. El espesor máximo de silicio que se puede examinar para una fuente de 200KeV es de 1,5 μm y de solo 0,5 μm para energía de 80KeV.

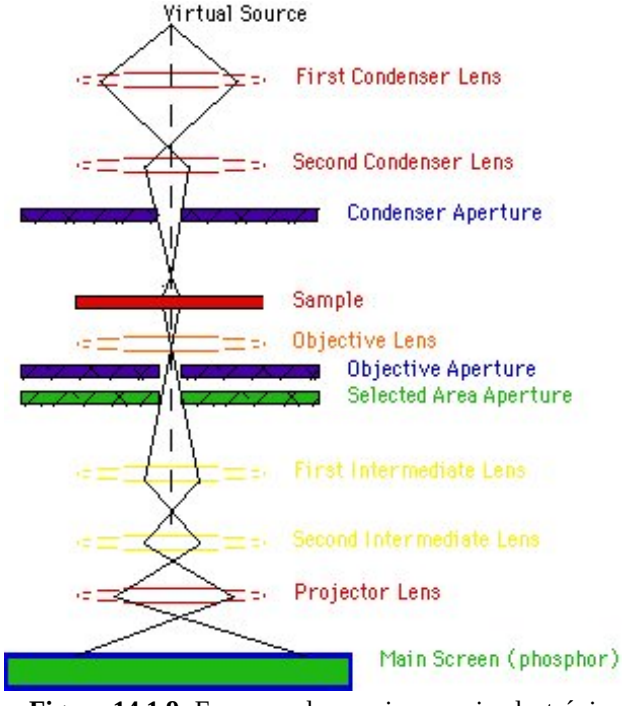


Figura 14.1.9: Esquema de un microscopio electrónico de transmisión
http://www.unl.edu/CMRAcfem/temoptic.htm

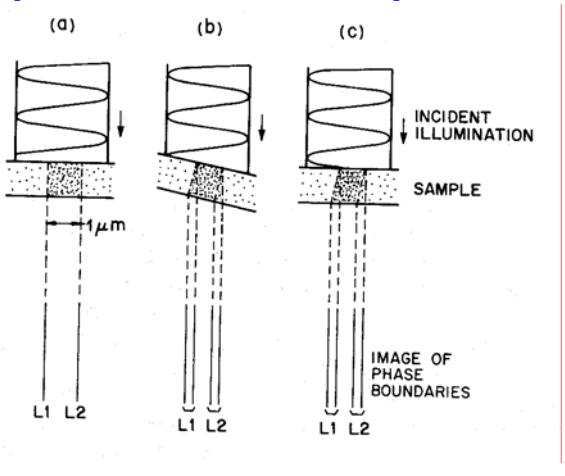


Figura 14.1.10: Análisis mediante TEM: a) de una muestra bien orientada con dos interfaces. b) de una muestra mal orientada. c) de una muestra bien orientada pero mal cortada.

Los especialistas en ULSI están normalmente interesados en las características morfológicas cuyas fronteras de fase se extienden de una cara a la otra de la muestra. Estas fronteras además limitan el espesor máximo de la muestra. Por ejemplo una muestra de delgada bien orientada que contiene una pista de polisilicio con una anchura de una micra sobre oxido produce una imagen de dos líneas. Correspondiente a cada una de las dos interfaces. Como se puede ver en la **Figura 14.1.10** (a).

Hay que tener un cuidado especial en la preparación y colocación de la muestra para el análisis con TEM. En la **Figura 14.1.10** podemos ver un dibujo de la sección de un corte vertical de un chip (el plano del dibujo es perpendicular a la superficie del chip). Si la muestra esta mal orientada se produce un desdoblamiento de las líneas que marcan los limites de las interfaces (b). Esto es un gran problema que aparece también si el corte de la muestra no se hace debidamente como se ve en el dibujo (c)

El contraste en un TEM es diferente según se trate de estructuras cristalinas o amorfas.

Para estructuras cristalinas el haz de electrones sufre una difracción por el material y, las variaciones en la intensidad de difracción producen un contraste en una imagen proveniente del haz no difractado (zona brillante) o desde uno o mas haces difractados (zona oscura). Los cambios abruptos de espesor, fase de la estructura o orientación cristalográfica causan cambios abruptos en el contraste y estas características cristalográficas pueden ser vistas fácilmente con gran resolución

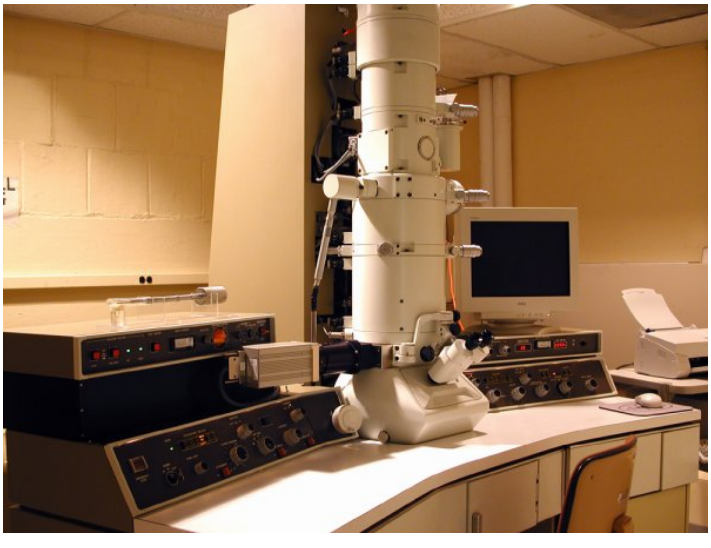


Figura 14.1.11: Microscopio Electrónico de Transmisión JEOL **200CX**
http://www.med.sc.edu:89/IRF/EQUIP.HTM
http://www.jeol.com/wi/wi.html

En el caso de materiales amorfos el contraste viene determinado por los cambios locales en la dispersión de los electrones debido a las diferencias de espesor, distinta composición química o diferencia de fase. Una región cuyo espesor varía continuamente produce una variación continua en la intensidad de la imagen, a diferencia del caso de contraste por difracción. Las imágenes obtenidas de Óxidos Nitratos y otros materiales amorfos son mas fáciles de interpretar que las imágenes de materiales cristalinos.

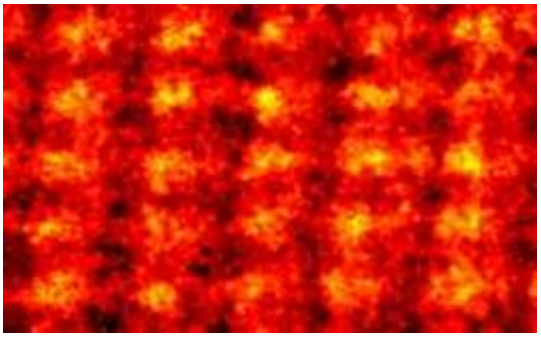


Figura 14.1.12: microfotografía de un cristal de silicio con una gran resolución (**http://focus.aps.org/story/v2/st24**)

La dificultad en la preparación de las muestras es el principal factor que limita la aplicación del TEM al estudio de ULSI. Una gran dificultad es la preparación de una muestra lo suficientemente delegada. Otra es que después de preparar la muestra las características morfológicas de interés deben estar presentes en la fina región

14.1.2 Análisis Químico

Se necesitan diferentes métodos para el análisis químico de materiales en la tecnología ULSI. La resolución espacial que se requiere va desde dimensiones atómicas, hasta dimensiones macroscópicas. Las resoluciones espaciales verticales y laterales son diferentes en estos estudios. Son necesarios diferentes requerimientos de sensibilidad desde 1011At/cm^3 hasta 1022at/cm^3. Las sustancias químicas buscadas en estos estudios son dopantes (arsénico , boro, fósforo), oxigeno , carbón, componentes de las metalizaciones, impurezas metálicas, es decir, una gama extensa de elementos desde los elementos ligeros hasta los pesados.

14.1.2.1 Auger Eelctron Spectroscopy (AES)

La espectroscopia Auger analiza una clase de electrones llamados electrones Auger. Estos se generan al incidir sobre una muestra un haz de electrones. Lo que produce que estos electrones Auger se desprendan de la muestra con diferentes energías dependiendo del tipo de material que se trate.

El mecanismo por el cual se liberan electrones Auger es el siguiente: Un electrón o un fotón con suficiente energía incide sobre la

muestra, provoca que un electrón de la capa K salga despedido, otro electrón del nivel L1 pasa a ocupar el nivel de energía K en este proceso se produce una energía que provoca que un electrón de la capa L2,3 , electrón Auger , salte al vacío.

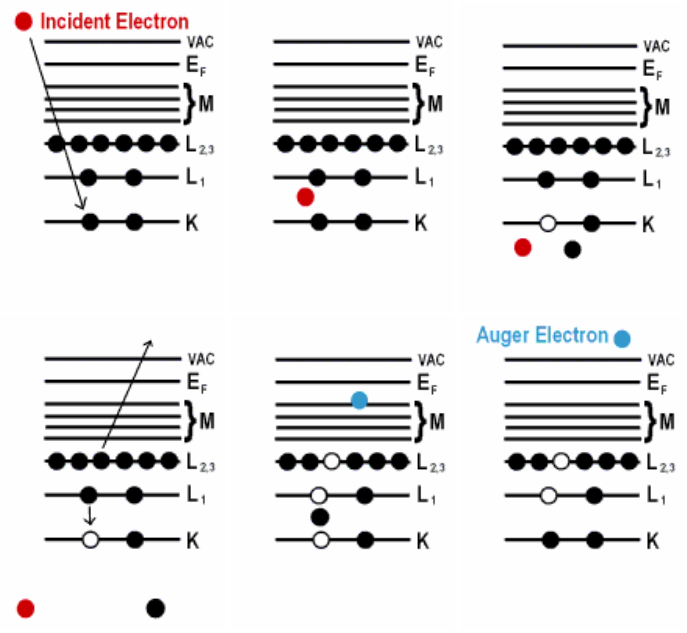


Figura 14.1.13: Generación de un electrón Auger (http://www.almaden.ibm.com/st/scientific_services/materials_analysis/)

La energía del electrón incidente esta comprendida entre los 2 y 10 KeV y penetra una pequeña distancia. Las energías de los electrones Auger van de los 20 a 2000Ev energías entre las bajas energías de los electrones secundarios y las altas energías de los electrones dispersados hacia atras (backscattered electrons). La energía de escape de los electrones Auger es menor de 50Å y mas pequeña cuanto menor es a energía de transmisión. Ver Tabla 14.1.3

Elemento	Energía de transmisión (eV)	Profundidad de escape (Å)
Fosforo	120 1859	5 32
Boro	179	6
Oxigeno	507	12
Arsénico	1228	23
Aluminio	1396	26
Silicio	92 1619	4 29

Tabla 14.1.3: Profundidad de escape para los electrones Auger

Se puede realizar por tanto un análisis químico de la superficie de una muestra usando AES.

Algunos diagnostico de problemas requieren un análisis mas profundo que la profundidad de escape de los electrones Auger. En ese caso la muestra es horadada por iones como en el Sputtering para crea una nueva superficie. Los datos son obtenidos cada cierto tiempo parando el proceso de horadado, o continuamente junto con el proceso de horadado. Las alturas de los picos Auger pueden ser representados en función del tiempo de horadado o de la profundidad y se puede obtener un pefil profundo de una muestra.

Para el análisis cuantitativo de la concentración de un elemento i Ci en el material base se utiliza la siguiente expresión

$$C_i = \frac{\alpha_i I_i}{\sum_j \alpha_j I_j} \qquad (14.2)$$

Donde Ii es la intensidad de pico Auger del elemento i y Ij es la intensidad de pico del elemento base. La constante de proporcionalidad α pude ser fácilmente determinada de estandares conocidos.

La resolución espacial lateral viene dada por el tamaño del haz incidente. Actualmente se consiguen resoluciones de hasta 20nm.

El inconveniente de esta técnica es que es destructiva.

Figura 14.1.14: JAMP-7810 Scanning Auger Microprobe

http://www.jeol.com/sa/saprods/jamp7810.html

14.1.2.2 Análisis por activación de neutrones (Neutron Activation Análisis) NAA

El análisis por activación de neutrones es el método analítico mas sensible para detectar la presencia de gran cantidad de elementos químicos. Con este método se puede determinar de forma precisa la concentración de elementos en una muestra de material. Tiene la suficiente sensibilidad para medir ciertos elementos a nivel de nano gramos e incluso por debajo.

Las ventajas de este método son su alta sensibilidad, la posibilidad de analizar a la vez de hasta 40 elementos y además es una técnica no destructiva

Únicamente no se pueden analizar los materiales ligeros como Boro Oxigeno Nitrogeno y Carbon porque no producen isótopos radioactivos analizables o tienen vidas medias muy cortas.

El método se basa en la detección y medida de los rayos gamma emitidos pro isótopos radioactivos producidos en una muestra sometida a una irradiación de neutrones.

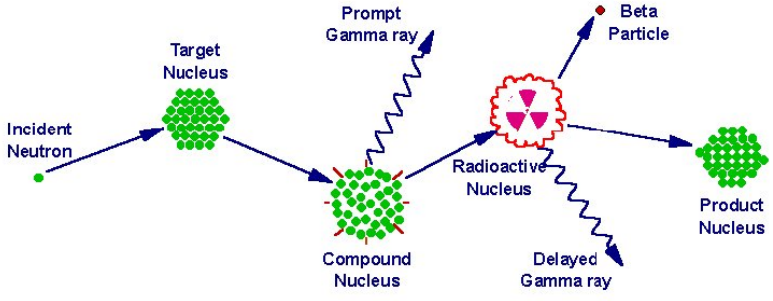


Figura 14.1.15: Esquema del proceso de captura de un neutrón por parte de un núcleo y posterior emisión de radiación gamma (http://www.missouri.edu/~glascock/naa_over.htm)

La irradiación se realiza con un flujo de neutrones térmicos de 10^{13} a 10^{14} /cm^2-s durante un periodo de 0.5 a 12 horas. Durante la irradiación de las obleas de silicio se produce un numero de especies radioactivas, incluyendo Si^{31} . La vida media del Si^{31} es de 2,6 horas. Después de la irradiación se deja reposar la muestra durante 24 a 48h para permitir que el nivel de radiación del silicio descienda por debajo de los niveles de otros elementos.

La radiación más comúnmente monitorizada durante el proceso de detección es radiación γ con energías entre los 0.1 y 2.5 MeV. Esta radiación es detectada por un detector de germanio y analizada por un analizador multicanal.

Ambas la energía de emisión γ y la medida de la vida media son utilizadas para la identificación del isótopo que aumenta la radiación. Para evaluar la cantidad de elemento presente varios factores deben conocerse como la cuenta de la radiación de un intervalo de tiempo, las eficiencias del emisor y detector para un pico particular, el flujo de neutrones térmicos, el tiempo de reposo desde la irradiación, el tiempo de irradiación etc.

El número de átomos en de un elemento particular en una muestra puede calcularse a partir de formulas conocidas y la correspondiente concentración en volumen puede ser calculada.

NAA es muy útil para medir la concentración de impurezas y para la resolución de problemas de contaminación introducidos en los procesos de fabricación. Pequeñas cantidades de impurezas pueden degradar la vida media de los portadores en las regiones activas de los C.I. Muchas impurezas tienen una actividad eléctrica significativa para concentraciones incluso de 1 ppmm (una parte por mil millones). Las medidas de las concentraciones de estos elementos en el semiconductor son esenciales para el desarrollo de los dispositivos y el mantenimiento del control de calidad durante su manufacturación.

14.1.2.3 Rutherford Backscattering Spectroscopy RBS

La Espectroscopia Rutherford RBS consiste en detectar la energía y flujo de los iones incidentes que son dispersados hacia tras (backscattered) mediante colisiones elásticas con los atomos de una la muestra.

En la técnica RBS se dirige un haz de iones con energías entre 1 y 3MeV hacia la superficie de la muestra. El diámetro del haz es de 10μm a 1mm. Los iones al incidir con los átomos de la superficie sufren choques elásticos que les hacen perder energía salen dispersados hacia tras (backscattered) y son detectados por un detector de barrera de silicio de energía dispersiva y procesada en un analizador multicanal. En RBS solo se detectan los iones dispersados hacia tras, y esto solo se produce si el átomo diana es mas pesado que el ión incidente. Normalmente se usan iones de He4 .

El factor cinético K relaciona la energía del ion incidente E_0 con la energia del ion reflejado E_0'

$$E_0' = K\, E_0 \qquad (14.3)$$

Ya que el valor K es conocido para cada elemento. La composición química en la superficie de la muestra puede ser determinada midiendo la energía de los iones reflejados

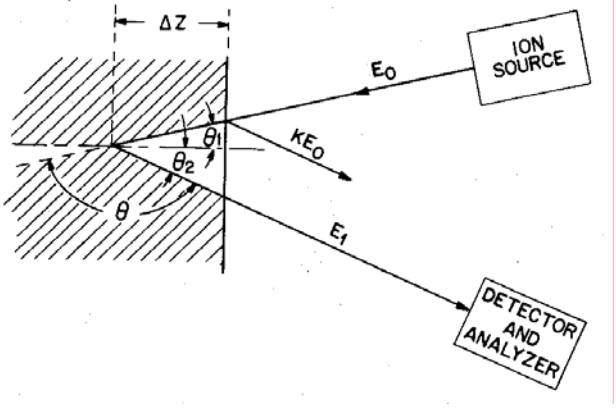


Figura 14.1.16: Esquema del análisis de una muestra por el metodo RBS.

Si los iones incidente penetran en la muestra antes de sufrir el choque entonces se produce una perdida de energía. Los iones dispersados hacia tras para una profundidad ΔZ que salen de la muestra sufren una perdida adicional de energía. La diferencia total de energía entre los iones dispersados en superficie y los que se introducen una distancia ΔZ viene dada por la expresión

$$\Delta = KE_0 - E_1 = [\varepsilon] N \Delta Z \qquad (14.4)$$

Donde $[\varepsilon]$ es el factor de sección de corte de parada y N es la densidad atómica de los elementos de la muestra. Se puede obtener un perfil de profundidad representando el numero de iones dispersados hacia tras en función de las energías de los iones E1

RBS es de los pocos análisis químicos que pueden proporcionar información cuantitativa sin el uso de estándares.

RBS no es una técnica destructiva ya que no existe daño de radiación, sputterin o transferencia de carga

14.1.2.4 Secondary Ion mass Spectroscopy SIMS

La espectroscopia de masa de los iones secundarios SIMS es un método destructivo de análisis de la superficie con una sensibilidad bastante alta. Este consiste en el bombardeo de iones de la muestra que provoca la extracción de iones secundarios que son analizados con un espectrómetro de masas.

Se utilizan iones positivos y negativos dependiendo del material que queremos estudiar con energías entre 5 y 15keV. Un haz de iones es rastreado a lo largo de una pequeña área de la superficie para crear un cráter poco profundo de fondo plano. El análisis de masa se realiza sobre la fracción de iones de material desprendidos de la parte central del cráter. Cuando se usan corriente de iones primarios muy bajas se realizara solo un análisis superficial de unas pocas capas atómicas. Para obtener perfiles mas profundos se usan corriente de iones primarios mayores.

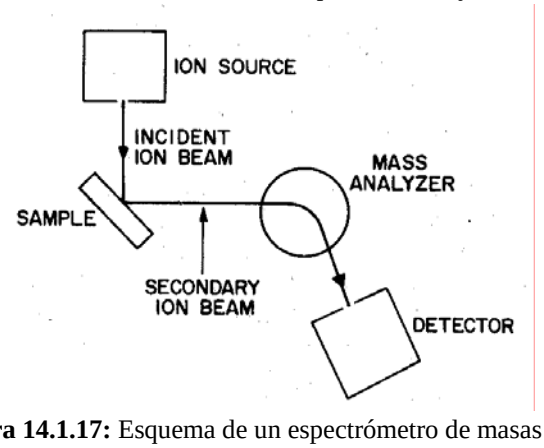


Figura 14.1.17: Esquema de un espectrómetro de masas de iones secundarios. “libro VLSI Technology de Sze”

La resolución espacial viene determinada por el tipo de iones utilizados, consiguiéndose resoluciones espaciales de 0.5μm. Sucede que a mayor resolución se pierde de sensibilidad. La sensibilidad es bastante grande de 1 ppm, en la siguiente tabla podemos ver las cantidades mínimas detectables para varios elementos en un substrato de Si.

SIMS se usa ampliamente para la caracterización de implantación iónica, perfiles de dopado, el análisis de laminas delgadas, análisis de trazos de contaminación,. Por tanto es una importante herramienta para industria microelectrónica en las áreas de control de calidad, análisis de fallos y desarrollo de procesos.

Elemento	Iones primarios	Iones secundarios	C_{min} (atomos/cm^3)
Arsénico	Cs^+	$^{75}As^-$	$5x10^{14}$
Fósforo	Cs^+	$^{31}P^{\pm}$	$5x10^{15}$
Boro	O_2^+, O^-	$^{11}B^+$	$1x10^{13}$
Oxigeno	Cs^+	$^{16}O^-$	$1x10^{17}$
Hidrogeno	Cs^+	$^{1}H^-$	$5x10^{18}$

Tabla 14.1.4: Parámetros SIMS para la detección de algunos elementos relacionados con problemas de la tecnología

14.1.2.5 Espectroscopia de emisión de rayos X (XES)

Cuando comentamos la técnica SEM vimos que al incidir un haz de electrones sobre una muestra se pueden generar rayos X. En la técnica XES se analizan esta emisión de rayos X. Cada material tiene un espectro característico de emisión de rayos X con picos identificables característicos.

La resolución espacial depende del volumen de la muestra que contribuye a la emisión de rayos X. Estos son generados durante la dispersión del haz de electrones incidentes y de los electrones dispersados hacia atrás (backscattered).

La sensibilidad y la precisión en el análisis cuantitativo viene determinada por la eficiencia de la generación de rayos X, la interferencia con otros picos y con el fondo de radiación y por el detector y otros parámetros de los instrumentos. En condiciones optimas se pueden conseguir sensibilidades de 10-15 gramos sobre un substrato de otro material. Lo que correspondería a una partícula de oro de 100Å de diámetro.

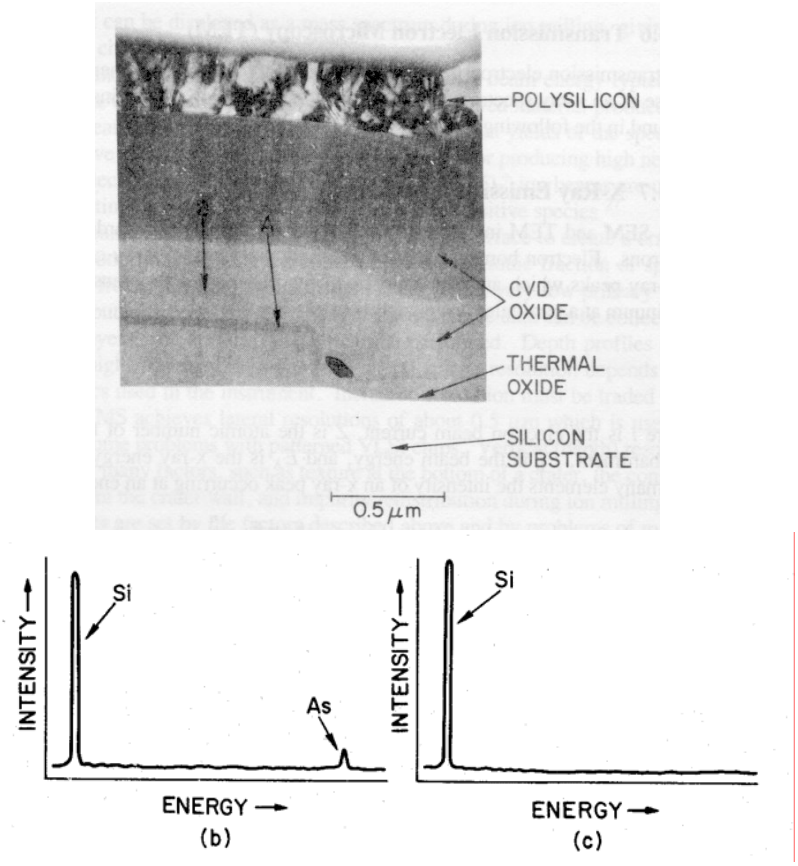
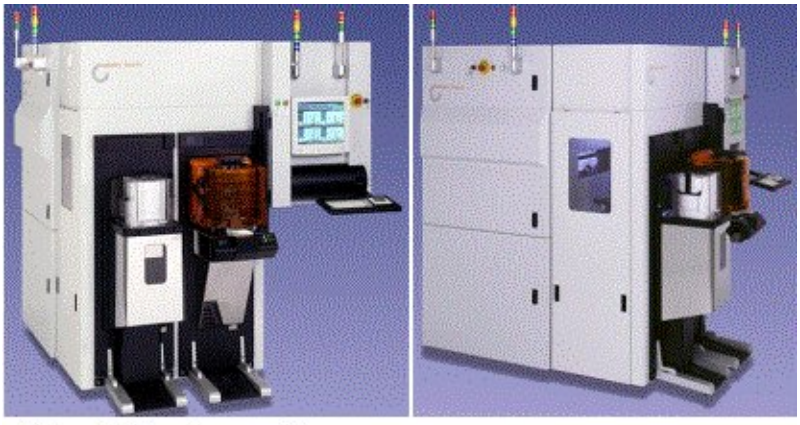


Figura 14.1.18: Sección de un NMOS. (b): análisis XES de una banda de oscura de 300Å de ancho. (región A) la figura (b) muestra el pico correspondiente a la presencia de arsénico. (c) análisis de de la región de substrato sin arsénico.

14.1.2.6 X-Ray Fluorescense (XRF)

En este caso se estimula la muestra con una radiación primaria de rayos X y se analiza la emisión de rayos X de la muestra bombardeada. Del espectro de radiación emitida por la muestra se puede realizar un análisis cualitativo cuantitativo. Los métodos usados para la detección del espectro de rayos X son parecidos a los usados en la técnica XES.



300mm Bridge System with one SMIF and One FOUP load port

Side View of Bridge System

Figura 14.1.19: Analizador por fluorescencia de rayos X de la casa Jordan Valley
http://www.jordanvalleysemi.com/product.asp
http://www.sea.rl.ac.uk/newsea/newpubs/combimexx/wyon.pdf

XRF se suele utilizar para analizar muestras de material aislante como óxidos y polímeros usados en el empaquetamiento. Estos materiales son difíciles de examinar con la técnica XES ya que se cargan negativamente o se descomponen durante el bombardeo de electrones. Sin embargo con la técnica XRF son fácilmente analizables. La gran anchura del haz de rayos X incidentes hace que este análisis no sea apropiado para pequeñas regiones del chip. Pero es muy útil para áreas mayores. Ya que los rayos X penetran a mayor profundidad que los electrones XRF será útil para el estudio de estructuras multicapas.

14.1.2.7 XPS ESCA

La espectroscopía de fotoelectrones por rayos X (X-ray Photoelectron Spectroscopy) XPS también llamada análisis químico por espectroscopia de electrones (ESCA) se usa en la industria para investigar la composición de los materiales y sus propiedades electrónicas como anchura de bandas o gaps.

Esta es una técnica que consiste en el bombardeo de rayos X de una muestra y posterior análisis de la energía de los fotoelectrones emitido por ella. La diferencia entre la energía de los rayos X y las energías de los fotoelectrones nos da las energías de los enlaces de los electrones del nivel más próximo al núcleo.

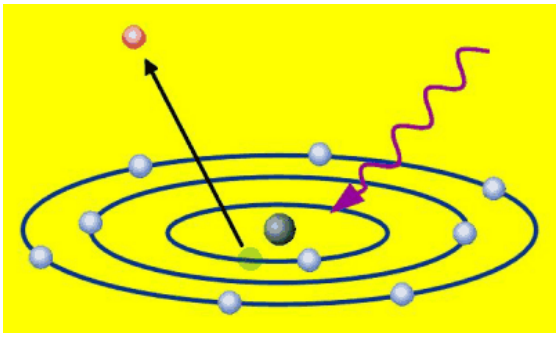


Figura 14.1.20: Haz de rayos X incidiendo sobre un átomo y posterior expulsión de un electrón de su capa fundamental

http://www.almaden.ibm.com/st/scientific_services/materials_analysis/xps/

La detección de los electrones y la instrumentación utilizada en parecida a la que se hace en el método AES. La resolución en profundidad es parecida en ambos métodos. La resolución espacial lateral de XPS es bastante pobre ya que el diámetro del haz de rayos X es bastante grande.

Sin embargo XPS tiene la ventaja de ser un método no destructivo. Además puede ser aplicada para el estudio de materiales aislantes ya que no tiene el problema de cargas en la superficie al usar un haz incidente neutro.

14.1.3 Estudio de las Propiedades mecánicas y Cristalográficas.

Un aspecto importante de los materiales de los dispositivos y los programas de desarrollo de los procesos es el análisis de las propiedades cristalográfico y mecánico de las capas y substratos del C.I. Estos análisis incluyen

a) La determinación del a orientación del sustrato

b) El grado de la orientación preferente

Los métodos utilizados en estos análisis son los siguientes:

1) Reflectancia Láser usado para medir la curvatura de la oblea para saber el estrés. LR

2) Rutherford backscattering Spectroscopy (Channeling). RBS(Channeling)

3) X-ray Difraction (metodos de camara y difractometro). XRD

4) Transmisión Electrón Difracttion. TED

5) Transmisión electrón Microscopy. TEM

En la **Tabla 14.1.5**se puede ver el campo de aplicación y su eficacia de los diferentes metodos para los distintos problemas de la estructura

Tipos de problemas	XRD Camera	XRD Diffrac-tometer	TED	RBS (channe lling)	LR	TEM
Identificación de fase	Bueno	Bueno	Bueno	No	No	No
Orientación preferente	Bueno	Pobre	Bueno	No	No	No
Parámetro celda unidad	Aceptable	Bueno	Aceptab	No	No	No
Presencia de regiones amorfas	Bueno	Bueno	Bueno	Bueno	No	Bueno
Localización en la red de átomos de impurezas	No	No	No	Bueno	No	No
Orientación del subtrato	Bueno	Aceptable	No	No	No	Pobre
Análisis de defectos cristalograficos	No	No	No	No	No	Bueno
Stres	Pobre	Bueno	Pobre	No	Bueno	Pobre

Tabla 14.1.5: Campo de aplicación de varios métodos analíticos para la problemas relacionados con la estructura

14.1.3.1 Laser Reflectance (LR)

Con este método se puede medir la curvatura de la oblea y calcular el estrés de películas delgadas sobre substratos gruesos. El método consiste en medir la diferencia de posición de un haz reflejado en la oblea. El haz reflejado es recogido a una distancia grande (10m). La oblea se mueve una distancia conocida mientras es iluminada por el láser. La curvatura de la oblea causa un desplazamiento en el haz reflejado. Para medir el estrés de una película delgada se mide la curvatura antes y después de depositarla.

Figura 14.1.21: Sensor Óptico Multihaz montado en un sistema de deposición de película delgada comercial.

http://www.moorfield.co.uk/pr01.htm

14.1.3.2 Rutherford backscattering Spectroscopy (Channeling)

La espectroscopia Ruterford además de usarse para analizar la composición química de un material se utiliza para la determinación de la estructura cristalina y los defectos característicos de un cristal.

Si la dirección del haz de iones coincide con un eje cristalográfico de bajo índice o con un plano de alta simetría se produce el efecto "channeling". Los iones penetran profundamente en el material sin encontrar ningún obstáculo. Si se cambia el ángulo de incidencia ligeramente la dispersión hacia tras crece abruptamente.

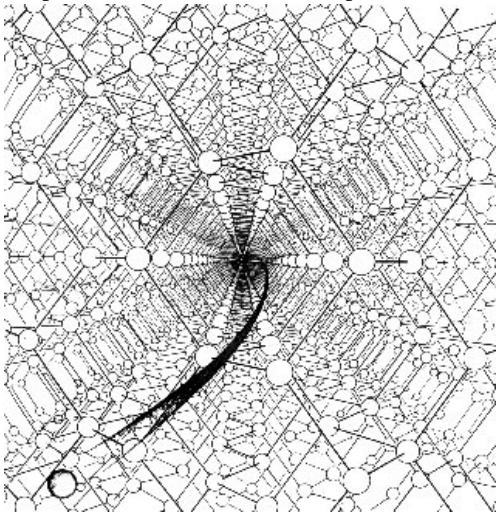


Figura 14.1.22: Dibujo de un electrón penetrando profundamente en un cristal debido al efecto Channeling
http://ionbeam.kigam.re.kr/oldpage/channeling_k.html
http://www.uam.es/otroscentros/cmam/espanol/tecnicas/node4.html

Para una orientación alineada se podrán detectar las anomalías en la perfección cristalográfica tales como la originadas por átomos intersticiales o defectos lineales ya que estas causan un incremento de los iones dispersados hacia tras. En el caso extremos de capas de material amorfas el rendimiento de los iones reflejados es idéntico a que se obtiene con una orientación aleatoria de un cristal perfecto.

Con este método también se puede determinar la distribución de dopantes entre lugares intersticiales y substitucionales.

14.1.3.3 Difracción de rayos X

El método de la difracción de rayos x consiste en el análisis de la posición angular e intensidad del haz de rayos X difractado por un cristal el cual proporciona información de la estructura cristalográfica del material.

Figura 14.1.23: Difractometro de Rayos X D8 Discover de la empresa Bruker AXS
http://www.bruker-axs.de/

Existen cuatro técnicas diferentes de difracción de rayos X :

1) La técnica Lau de “back reflection”
2) La técnica de cámara de lectura (Read Camera Technique)
3) La técnica de cámara de Huber-Seemann-Bohlin
4) El método de difractometro

Todos estos métodos se basan en el establecimiento de las condiciones que satisfacen el requerimiento de Bragg para la difracción de los rayos X por una red cristalina:

$$n\lambda = 2dsen\theta \qquad (14.5)$$

Donde λ es la longitud de onda del rayo X, d es el espacio interplanar, θ es el ángulo de difracción Bragg y “n“es un numero entero que indica el orden de la difracción. La difracción ocurre solo cuando se satisface la ecuación (14.5)

14.1.3.4 Transmision Electrón difraction (TED)

La técnica de Difracción por transmisión de electrones es parecida a la difracción con rayos X pero en este caso el haz incidente que atraviesa la muestra es de electrones. Igualmente tiene que cumplirse la ecuación para que se produzca en este caso la difracción de los electrones. Para realizar este análisis se usa un microscopio electrónico de transmisión. La imagen que se obtiene de un material policristalino es una seria de círculos concéntricos, en caso de un material cristalino una serie de puntos. La técnica TED se puede usar para la determinación de la orientación de pequeños cristales de Silicio que se producen dentro del oxido de polisilicio en el proceso de oxidación térmica del polisilico.

Figura 14.1.24: Imagen TED de una muestra delgada de oro policristalino. Cada anillo de difracción es debido a un plano especifico de átomos. ([100] o [111]). **http://www.world-mysteries.com/ sci_mpbpp.htm**

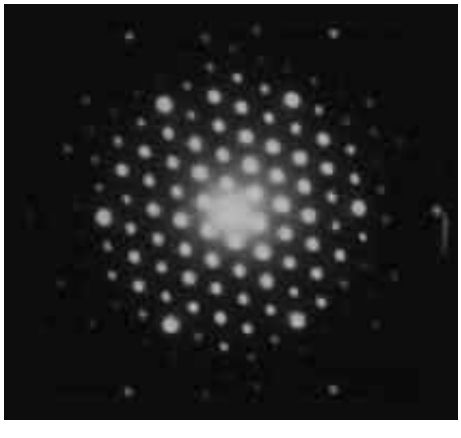


Figura 14.1.25: **Imagen TED de un material cristalino**

http://www.matter.org.uk/diffraction/electron/electron_diffraction.htm

14.1.4 Mapeo eléctrico

El mapeo eléctrico es un método de análisis que utiliza un haz de electrones para localizar regiones dentro de la estructura de un dispositivo con diferente actividad eléctrica y en algunos casos medir esta diferencia. Se suelen utilizar dos técnicas diferentes de mapeo eléctrico: la microscopia de contraste de voltaje y la microscopia de recuento de carga o EBIC (Electron Beam Induced Current). La primera se basa en la influencia del potencial de superficie local sobre la energía de los electrones secundarios que se producen cuando un haz de electrones incide en la muestra del dispositivo. El flujo de electrones secundarios que alcanza el detector proporciona por tanto información sobre ese potencial. Este fenómeno que es esencialmente una imagen de contraste de voltaje, puede utilizarse para determinar el potencial de un elemento sobre la superficie del dispositivo. La otra técnica se basa en la generación de cargas en el dispositivo producida por un haz de electrones y posterior recuento de esas cargas mediante una capacidad o unión. Los cambios locales en la morfología, propiedades de los materiales y el campo eléctrico de unión causan la modulación local correspondiente de la corriente registrada. Estas dos clases de mapeo eléctrico pueden realizarse con la ayuda de un microscopio de barrido. En el caso de del mapeo EBIC se puede utilizar un microscopio electrónico de transmisión.

14.1.4.1 Microscopia de contraste de voltaje

Para realizar la técnica de contraste de voltaje se utiliza un microscopio electrónico de barrido SEM. La imagen de los electrones secundarios se crea mediante un campo extractor de varios cientos de voltios que dirigen los electrones hacia un detector donde un alto campo eléctrico los atrae hacia una superficie brillante. Un tubo de luz transfiere la señal a un tubo fotomultiplicador. No todos los electrones alcanzan la superficie brillante. Los campos eléctricos locales de la superficie de la muestra afectan la trayectoria a de los electrones secundarios y permiten que solo sean detectados electrones con energías superiores a una umbral 1eV. El flujo total de electrones secundarios que alcanzan el detector y por tanto la intensidad de la señal depende del numero de electrones secundarios con energía mayor que la umbral. Este numero esta afectado por el potencial en a superficie como se muestra en la **Figura 14.1.26**. La mínima diferencia de intensidad que se puede apreciar en la pantalla corresponde a un potencial de superficie de un voltio. Lo cual se considera el limite practico.

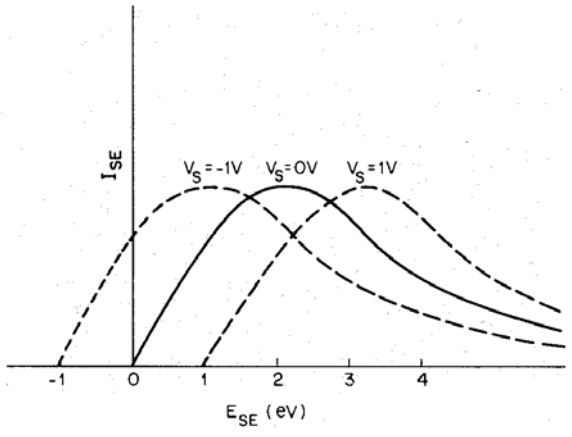


Figura 14.1.26: Intesidad de la señal debido a la emisión de electrones secundarios en funcion del potencial de superficie.

Esta técnica se puede aplicar para detectar las discontinuidades eléctricas en regiones conductoras de un circuito integrado. Para medir la discontinuidad eléctrica se aplica una tensión continua a una metalización. Esta polarización puede ser aplicada utilizando una punta o sonda de contacto móvil dentro de la cámara del microscopio de barrido o por un contacto fijo a los terminales de entrada en el dispositivo empaquetado.

Se puede también aplicar un pulso de tensión, por tanto, el dispositivo operaria de forma dinámica dentro del microscopio SEM y la imagen de contraste de voltaje obtenida mediante un pulso estroboscopico del haz de electrones incidente en sincronismo con el circuito. De esta forma en el modo de contraste de voltaje estroboscópico se pueden conseguir resoluciones de tiempo de 0.2 pico segundos.

Los circuitos de testeo, análisis y verificación de diseño a menudo requieren la caracterización de las formas de ondas en determinados nodos internos del circuito. Estas medidas son normalmente realizadas poniendo puntas de finas de prueba en esos nodos o en los terminales conectados a ellos. Como la anchura de las líneas de metalización es muy pequeña (aproximadamente 1micra), el hecho de situar la punta de prueba en el lugar seleccionado evitando dañar la metalización es bastante difícil. Un segundo problema con las puntas de contacto se debe a la capacidad de esta, generalmente mayor de 0.1pF, esta capacidad parásita en el nodo que se testea puede tener efectos sobre la mediada. Estos problemas se pueden resolver utilizando un haz de electrones de pequeño diámetro como sonda, el haz no añade ninguna capacidad en el circuito y puede ser posicionado fácilmente en elementos pequeños del circuito.

14.1.4.2 Microscopia de recuento de carga EBIC

La microscopia de recuento de carga o corriente inducida por haz de electrones EBIC se basa en el hecho de que al incidir un haz de electrones penetrando en un dispositivo se pueden generar portadores que pueden ser detectados en forma de corriente, la cual se puede monitorizar. Esta forma de análisis es útil para localizar fallos en los dispositivos dentro de una unión o una capacidad.

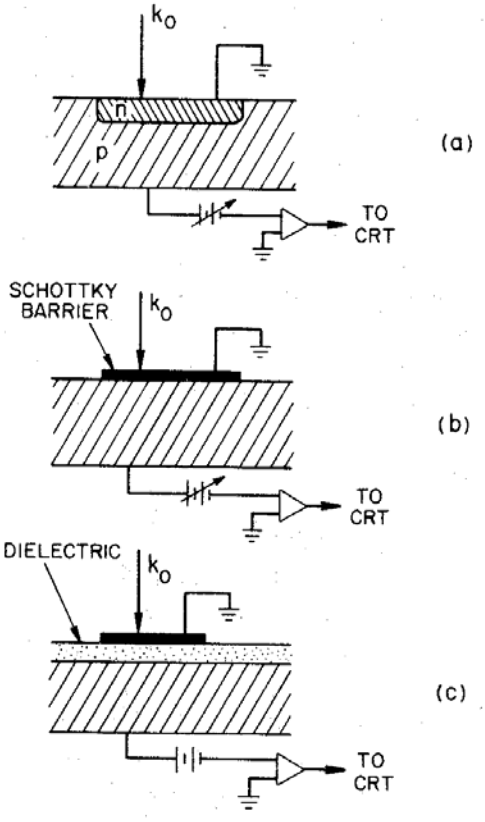


Figura 14.1.27: Esquema de la medida EBIC (a) de una unión PN, (b) de una unión de barrera Schottkly, (c) de una capacidad. La energía del haz incidente es k_0.

La técnica EBIC se suele utilizar para estudiar las barreras Schottky y las uniones PN detectándose los defectos en el silicio como las fault stacking, dislocaciones e inhomogeneidades segregadas durante el proceso de crecimiento del cristal. La microscopia EBIC puede utilizarse también para detectar defectos en los óxidos delgados de las capacidades debido al aumento local de la corriente túnel con un alto campo que aparece en esos sitios. En todos los casos los defectos crean una perturbación local de la corriente la cual crea un contraste de imagen. Por

tanto con el análisis EBIC se pueden localizar espacialmente los defectos responsables de las perdidas en las capacidades y su rotura.

En la **Figura 14.1.27** aparece un esquema de cómo se realiza el análisis EBIC para una barrera Schottkly una unión PN y una capacidad. Mediante una punta de contacto dentro de la cámara de un microscopio SEM se aplica una tensión. Se hace incidir un haz de electrones con una energía adecuada y una corriente de unos 10nA, generándose pares electrón hueco dentro o por debajo de la región de carga espacial que se difunden hacia la región de carga espacial y son recogidas dando una corriente. Los centros de recombinación debidos a los defectos que están en el camino que recorren los portadores generados disminuyen la corriente localmente que se recoge, formándose, de esta forma, una imagen de contraste con la cual se localizan estos defectos.

14.2 Técnicas de ensamblaje y empaquetamiento

El empaquetamiento del chip es un tema amplio que abarca desde la preparación de la oblea para el ensamblado hasta las técnicas de fabricación de empaquetamiento. El propósito del empaquetamiento del chip es proveer al chip de una conexión eléctrica, (expandir los "pitch" electrodos del chip para el siguiente nivel de empaquetamiento), proteger el chip de estreses mecánico y medioambientales y proporcionar un camino adecuado para la disipación del calor que se produce en el chip.

Después de ver todos los procesos planares de fabricación de un chip en muchos de los cuales se utilizan sofisticadas técnicas. Se podría pensar que el proceso de empaquetamiento es un proceso sencillo. Esto no es cierto el empaquetamiento requiere de las técnicas mas sofisticadas de diseño y fabricación, además, esta fase final es crucial en el buen funcionamiento de un circuito integrado y es la mas costosa económicamente de todo el proceso de fabricación.

Las fases por orden de fabricación en las que se puede dividir este proceso son las siguientes:

1. Preparación de la oblea
2. Interconexión del chip
3. Encapsulado
4. Testeo de fiabilidad

14.2.1 Tipos de empaquetamiento

En la tecnología ULSI se usan una gran cantidad de tipos de empaquetamiento. Estos se pueden dividir en dos tipos básicos: Los empaquetamientos herméticos cerámicos y los empaquetamientos plásticos.

En los empaquetamientos cerámicos el chip esta aislado del exterior mediante un cerramiento fuertemente al vació. Este tipo de empaquetamiento esta dirigido a aplicaciones de altas prestaciones donde puede ser asumido el alto coste de empaquetado.

En los empaquetamientos plásticos el chip no esta perfectamente aislado del exterior ya que el encapsulado esta formado por materiales de resina, normalmente resinas de tipo pegamento. El ambiente exterior afecta al chip pasado un determinado tiempo y gradualmente penetra en el plástico. Los empaquetamientos plásticos son más populares por su expansión de sus aplicaciones y sus mejoras de funcionamiento debido a los adelantos en los materiales plásticos y otros procesos de desarrollo. Debido a que su producción esta totalmente automatizada tienen un conste muy competitivo.

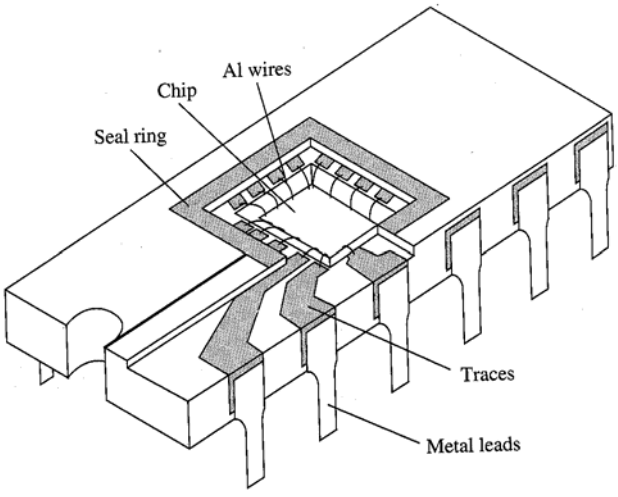


Figura 14.2.1: Dibujo de un empaquetamiento cerámico.

En la **Figura 14.2.1** podemos ver un empaquetamiento hermético cerámico típico. El chip se ubica en una cavidad del empaquetado. El material base del empaquetado es un molde cerámico sobre el cual se emplaza el cableado metálico y las patillas externas. El chip y el empaquetamiento esta interconectado por hilos finos de oro. El sellado hermético se completa con una tapa normalmente de material cerámico o de metal. El empaquetamiento evita los contaminantes externos y apenas tiene efectos mecánico ni químicos sobre el chip ya que los componentes del empaquetado no tocan la superficie del chip. El material usado para el empaquetamiento cerámico suele ser el Al_2O_3 también se usa el AlN cuando se requiere un gran capacidad de disipación.

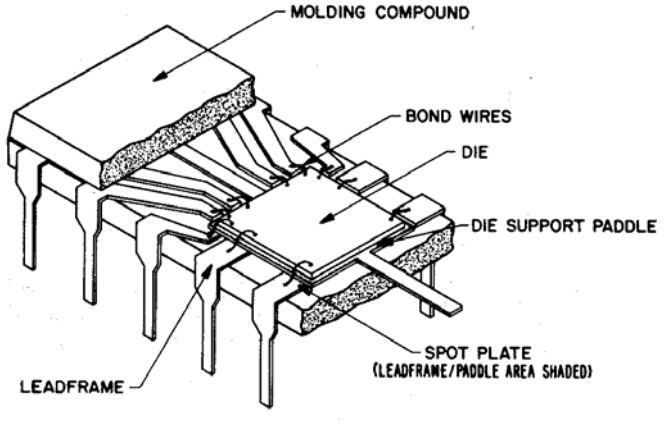


Figura 14.2.2: Esquema de un empaquetado plástico.

En la Figura 14.2.2 podemos ver un empaquetado plástico típico. En este caso el chip esta sujeto a la paleta de la estructura principal. La estructura grabada o estampada con una fina capa de metal sirve como esqueleto alrededor del cual se ensambla el empaquetamiento y

proporciona las patas externas del empaquetamiento completo. Las interconexiones se realizan con cables finos de oro. El encapsulado se extrae por el método de transferencia de molde utilizando un pegamento plástico. La resina cubre al chip y da la forma exterior del empaquetado al mismo tiempo. El alargamiento de las patillas externas se realizan después del moldeo. Este empaquetado no puede desacoplar el chip del ambiente exterior así que los efectos mecánicos y químicos sobre el plástico tendrán que tenerse en cuenta.

A parte de los encapsulados cerámicos y plástico. Los empaquetamientos se pueden clasificar según el nivel de integración siguiente, el nivel de placa impresa PWB (printed wiring board). Se tienen dos grandes grupos de empaquetamientos de un solo chip, los de atravesado por agujero TH (through hole) y los montados en superficie SM (sourface mount). Dentro de los TH se incluyen los DIP (dual in line package) y los PGA (pin grid array). Ambos están disponibles en encapsulados cerámicos y plásticos. Los empaquetamiento TH proporcionan una colocación mas sencilla de la pastilla para su soldadura y unas soldaduras mas fuertes con la placa impresa. Pero restan flexibilidad en el diseño de la placa impresa restringiendo la densidad de montaje.

El empaquetado de superficie SM (sourface mount) es el mas usado en la industria, sobre un 50% de todos los empaquetados son de este tipo y se utilizan en el empaquetado ULSI. Aunque este puede ser cerámico o plástico, se prefiere el plástico debido a su bajo coste de fabricación . Los empaquetados SM mejoran la flexibilidad de diseño de la placa impresa PWB y el aprovechamiento del espacio. Los empaquetados SM pueden tener dos tipos diferentes de formas de patillas las patillas J y las patillas de ala de gaviota (gull-wing). Los empaquetados SOJ (small-outline J-lead) y los QFJ (quad flat J-lead) son del tipo J y los empaquetados SOP (small-outline package) y los QFP (quad flat package) son del segundo tipo. Ver **Figura 14.2.3**

Para las aplicaciones ULSI de memoria donde se requieren pequeño numero de patillas, se utilizan los empaquetados J. Para los microprocesadores donde se necesitan una gran cantidad de patillas se suele usar los empaquetados de ala de gaviota.

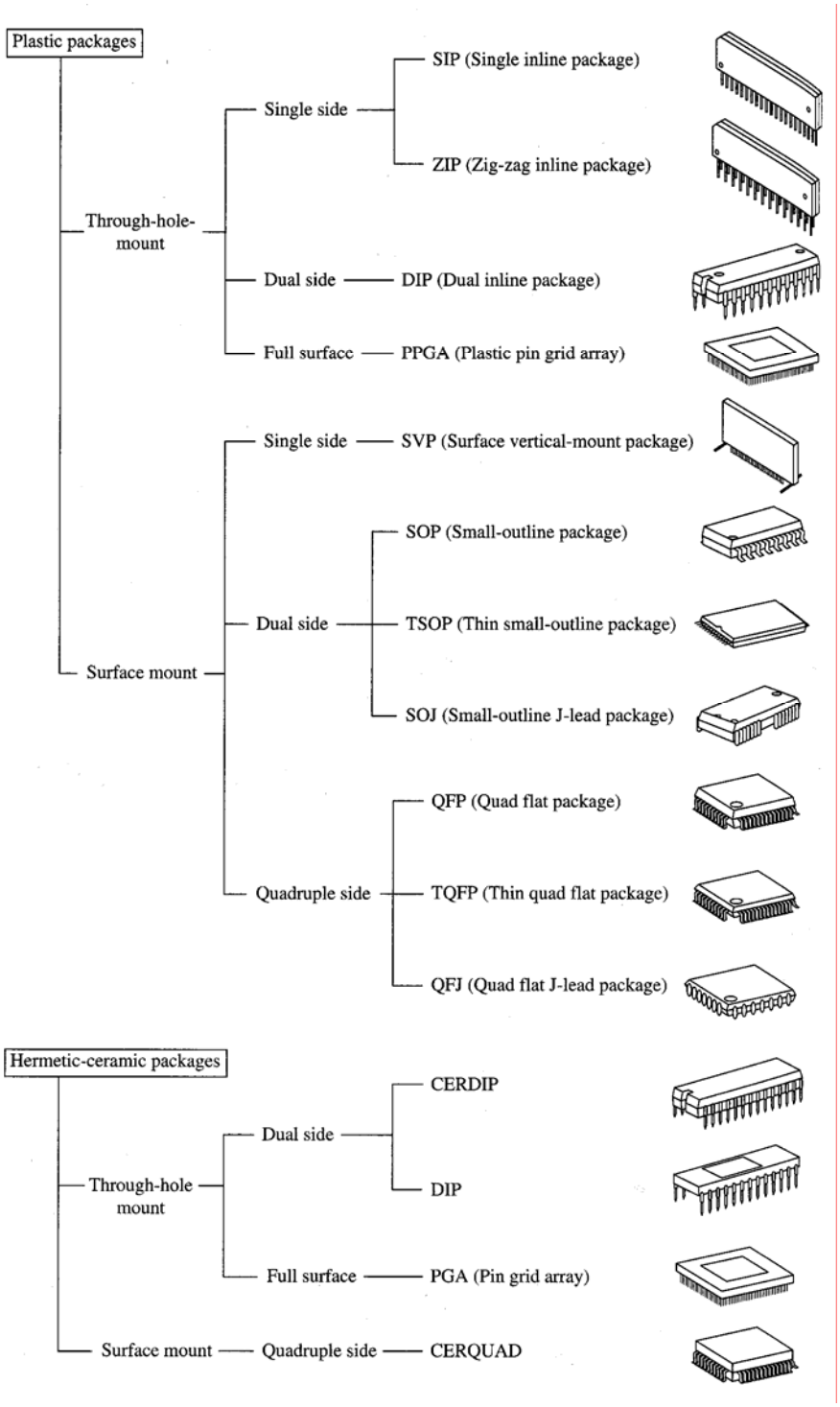


Figura 14.2.3: Clasificación y dibujo de los diferentes tipos de empaquetados del mercado.

14.2.2 Preparación de las obleas

El proceso de preparación de las obleas comienza con el pulido de la parte posterior para reducir el espesor de la misma (backgrinding). En los empaquetamientos montados en superficie (Surface Mount) los más usados en ULSI, un espesor delgado de la oblea es especialmente beneficioso para reducir el estrés térmico debido a la diferencia del coeficiente térmico de expansión (TCE) entre el chip y el material de la moldura de plástico. El proceso de pulido se realiza en maquinas automatizadas. Este debe ser cuidadoso. Una mala operación degrada la fortaleza mecánica de los chips y puede provocar roturas en los chips después de ser empaquetados.

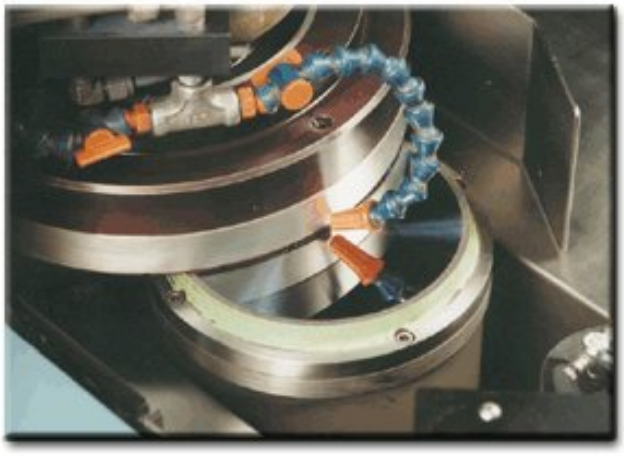


Figura 14.2.4: Maquina para le pulido posterior de la oblea (backgrinding) **http://www.icservice.com/**

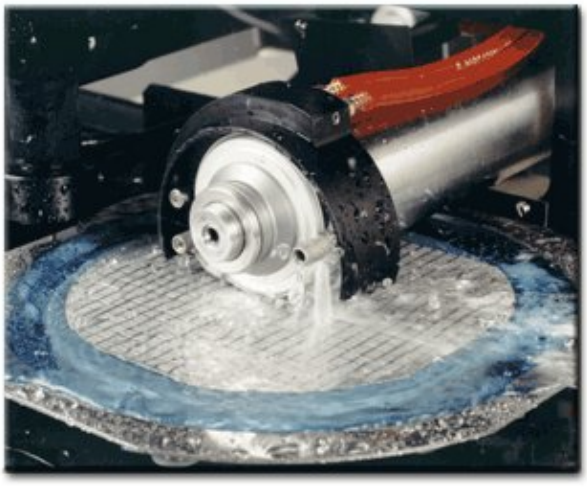


Figura 14.2.5: Proceso de cortado de la oblea (wafer dicing). **http://www.icservice.com/**

Después del pulido posterior (en caso de usar un pegado eutectico del chip) se realiza una metalización posterior de la oblea. Las metalizaciones mas usadas son de oro y metalizaciones multicapa de oro-níquel-plata o de Ti-Ni-Au. La combinación adecuada de los materiales de la mentalización posterior y las operaciones de pegado del chip proporcionan una mayor fortaleza de adhesión y conexiones eléctricas. Finalmente se procede a la obtención de los chips individuales cortando la oblea **Figura 14.2.5**. Las obleas se pegan con adhesivo a una cinta que ha sido preensamblada es un bastidor. Para el cortado se usa una sierra de diamante de 25 micras de espesor y que gira 20.000rpm. Esta corta la oblea a una profundidad que varia entre el 90 y 100% del espesor de esta. El corte con esta sierra permite secciones de corte de solo 60micras de ancho.

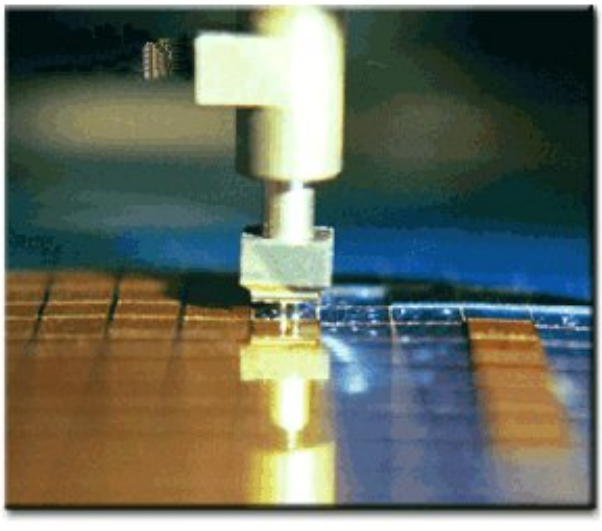


Figura 14.2.6: Proceso de selección y ordenación de los chips. (Die sorting). **http://www.icservice.com/**

Las maquinas para recortar los chips de la oblea son completamente automáticas y pose un sistema de alineamiento, una unidad de limpieza de la oblea, un horno de secado y un buen sistema supervisor. Las obleas cortadas son entonces cargadas para la operación de pegado del chip donde una juntadora automáticamente pone en orden los chips en buen estado reconociendo los chips defectuosos marcados o leyendo los datos del mapa de posición del las operaciones de test de la oblea. Ver **Figura 14.2.6**.

Video: http://www.surftape.com/solutions2.html

14.2.3 Interconexión del chip

La interconexión del chip consta de dos pasos: el pegado del chip (chip Bonding) y la unión con cables (Wire bonding)

En el primer paso el pegado del chip (chip bonding) se pega la parte de debajo del chip mecánicamente a un material base apropiado que pude ser un substrato cerámico o un substrato metálico. El pegado del chip además de proporcionar el soporte adecuado al chip para su posterior conexionado también proporciona un camino térmico para la disipación de calor del mismo, y algunas veces las conexiones eléctricas que se realizan por debajo del chip. En el segundo paso el unión por cables (Wire bonding) se realiza la conexión de los terminales del chip a los del empaquetado. Para ello se suele usar hilos finos de metal de oro o aluminio.

14.2.3.1 Pegado del chip (Chip bonding)

El pegado del chip es la primera operación de empaquetamiento propiamente dicha. A medida que se ha ido aumentando el nivel de integración se ha complicado la tecnología del pegado del chip. A la hora de realizar el pegado del chip habrá que tenerse en cuenta varios factores como el estrés térmico causado por los diferentes coeficientes térmicos de expansión (TCE) entre el chip y el material de empaquetado, la disipación de potencia, capacidad de corriente, fiabilidad y costo.

La diferencia de coeficientes térmicos de expansión (TCE) entre el chip y el material de empaquetado provoca problemas de estrés que pueden producir que se despegue o que se rompa el chip. Por eso lo ideal es escoger materiales para empaquetado con TCE lo mas próximo al Silicio.

Existen dos tipos de pegado del chip pegado por soldadura y pegado con pegamento. El primer método se usa en los empaquetamientos herméticos cerámicos y el segundo en los empaquetamientos plásticos.

14.2.3.1.1 Pegado por soldadura

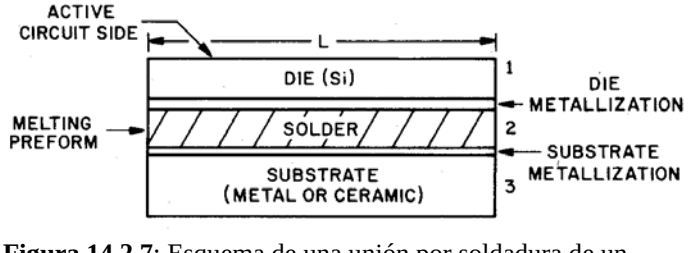


Figura 14.2.7: Esquema de una unión por soldadura de un chip la soporte del chip.

El pegado por soldadura esta libre de contaminación, tiene una excelente resistencia al desprendimiento, y asegura una baja humedad. La

mayor desventaja es la dificultad para automatizar el proceso y el estrés térmico al que se somete el chip debido a la alta temperatura del proceso.

En la figura 14.2.7 podemos ver como se realiza un pegado por soldadura. Tanto la parte de abajo del chip como la base del substrato tienen que tener una mentalización para que se pueda realizar bien la soldadura. El material que se usa para la soldadura es una lamina delgada de espesor menor de 0.05mm de una aleación adecuada (oro germanio por ejemplo). Al mismo tiempo que se aplica calor para realizar la soldadura se realiza un ligero movimiento o restregón del chip. El proceso se realiza por un brazo robotizado que coge el chip lo orienta y lo sitúa encima del substrato. La temperatura de fusión debe ser más elevada que la temperatura que se alcanza en el proceso posterior (wire bonding).

14.2.3.1.2 Pegado con pegamento

En el pegado con pegamento se usa una resina con trocitos de plata para mejorar la conducción eléctrica y térmica. Los pegamentos con plata son mucho mas baratos que las soldaduras con alto contenido en oro, además son mas flexibles en la absorción del estrés térmico entre el chip y el substrato.

El chip no necesita la mentalización posterior ya que ya que el pegamento proporciona una mejor adhesión con el dióxido de silicio de atrás, al substrato si se somete a una mentalización. En el proceso de pegado no se necesita un restregado, por tanto la probabilidad de daño de los bordes del chip es menor. El proceso de pegado se puede automatizar fácilmente. Con este método se consigue una velocidad de pegado e un chip por segundo. En el proceso de pegado se tendrá que completar con un tratado a altas temperaturas de 125 a 175ºC con una duración de una a dos horas.

En el pasado la resina tenia muchos iones hidrolizables, lo cual puede afectar directamente a fiabilidad. En la actualidad se usan resinas con concertaciones muy pequeñas de iones. Las resinas son ampliamente usadas en la actualidad en la tecnología ULSI.

14.2.3.2 Unión con cables (Wire bonding)

El proceso de unión con cables es el método mas utilizado para realizar la conexión de los terminales del chip con los terminales del encapsulado. Los cables que se utilizan son unos hilos finos de oro o de aluminio (de unos 20 a 30 micras de diámetro). Se eligen estos materiales porque con ellos se consigue una buena unión con los terminales del chip y las metalizaciones del empaquetado.

Para realizar la unión existen varias técnicas: termo-compresión, termosónico y vibración ultrasónica. Dependiendo de la forma de la unión se tiene dos tipos de unión: la unión de bola y la unión en cuña. En el caso del oro se suele usar la unión de bola por termo-compresión para los contactos del chip junto con una unión en cuña en los terminales del substrato del empaquetado, lo que se llamaría **la unión bola cuña** (ball-

wedge bonded). Cuando se usa el aluminio como material las dos uniones son de cuña y se realizan por vibración ultrasónica, **unión cuña-cuña**, (wedge-wedge bonded).

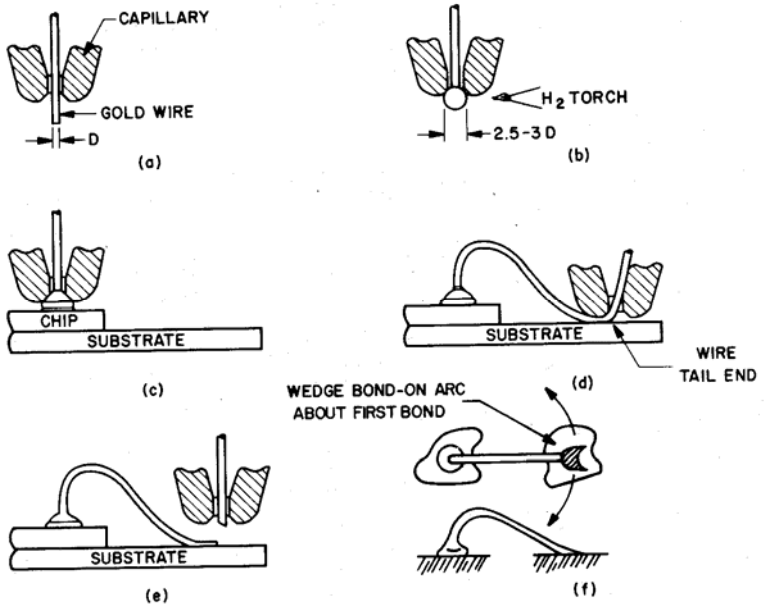


Figura 14.2.8: Esquema como se realiza una de una Union bola-cuña (ball-wedge bonded)

Figura 14.2.9: Fotografia del proceso de wire bonding del tipo una unión bola-cuña **http://www.netmotion.com/**

http://www.smallprecisiontools.com/caps_bond_cycle.htm

En la **Figura 14.2.8** podemos ver un esquema de como se realiza una unión bola-cuña usando hilos de oro. El proceso consta de los siguientes pasos:

Se monta una bobina de hilo fino de oro en el aparato de cableado, el cable de oro se hace pasar por un capilar de cristal o carburo de tungsteno (a). Una llama de gas de hidrógeno se hace incidir sobre el trozo de hilo de oro que sale por debajo del capilar formando una bola (b). Se baja el capilar hasta hacer incidir la bola de oro con el contacto del chip. El capilar ejerce la necesaria presión sobre la bola formándose la unión entre la bola de oro y el contacto de aluminio del chip (c), (en caso de que el sistema fuera termosónico se aplicaría después una vibración ultrasónica). Después se sube el capilar dejando que corra el cable a través de el y se lleva hasta el terminal del substrato metálico, se baja de nuevo el capilar hasta el terminal del empaquetado (d), mediante la combinación de temperatura y presión del capilar se forma una unión del cable al terminal (en una maquina termosónica se aplicaría luego una vibración ultrasónica), después se sube el capilar y se cierra a una altura programada ,el capilar sigue subiendo de forma que el cable se corta en la parte mas delgada de la unión (e). De esta forma se tiene nuevamente un trozo de hilo que sobresale por la parte inicial del capilar y se vuelve a repetir el proceso con la formación de la bola de oro mediante una llama de hidrógeno.

La ventaja del método de unión bola cuña es que después de realizar la unión de bola, la unión con el empaquetado se pude realizar en cualquier otra dirección mientras que en las uniones cuña-cuña las dos uniones tiene que estar alineadas. Por tanto la versatilidad del método de unión bola-cuña permite un mayor potencial de automatización. El método de unión bola-cuña con hilo de oro es ampliamente usado en la tecnología de empaquetamiento ULSI. En la actualidad las maquinas de cableado totalmente automatizadas del tipo unión bola-cuña tiene una velocidad de seis cables por segundo.

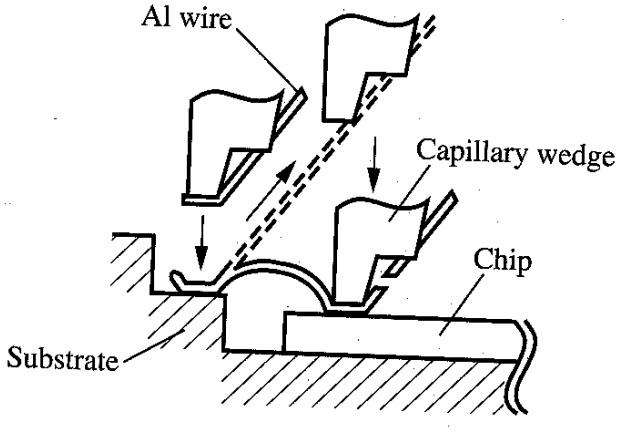


Figura 14.2.10: Union cuña-cuña, (wedge-wedge bonded)

En la **Figura 14.2.10** podemos ver un esquema de como se realiza una unión cuña-cuña usando hilos de aluminio. El proceso consta de los siguientes pasos:

Un capilar en forma de cuña portando el hilo de aluminio desciende y presiona el hilo contra el terminal en el substrato aplicando una vibración ultrasónica. Se levanta el capilar dejando el hilo libre y se lleva hacia el contacto del chip haciendo un arco, el capilar en forma de cuña vuelve a bajar y presiona el hilo contra el contacto del chip aplicando de nuevo una vibración ultrasónica. El capilar se levanta sujetando el hilo de forma que al tirar de el se rompe finalizando el ciclo de pegado. Las dos direcciones de las uniones deben ser iguales en este proceso

Las ventajas de usar hilos de aluminio es que la unión con los contactos de aluminio del chip son mejores desde el punto de vista metalúrgico. En las uniones Oro Aluminio se pueden producir compuestos intermetálicos que son frágiles y pueden producir la rotura de la unión. Este método de cableado del chip es normalmente utilizado en los empaquetamientos cerámicos.

14.2.4 Encapsulado

14.2.4.1 Empaquetamiento cerámico

Para realizar el empaquetado se utiliza una técnica multicapa. Partiendo de una disolución de material cerámico se transforma en una tira fina. Después de secarse las planchas se cortan a un determinado tamaño, se realizan mecánicamente una serie de agujeros de lado a lado (a través de los cuales se realizaran las interconexiones) y cavidades en las laminas. Se realizan el cableado rellenado las cavidades y agujeros con una disolución de tungsteno. Se ponen las laminas de forma alineada y son prensadas, posteriormente la estructura completa es sometida a una temperatura de 1600ºC para formar una estructura monolítica. Después se realizan realiza el patillado y chapado metálico. Finalmente, una vez que se ha colocado el chip y sus conexiones se coloca una tapa cerámica o metálica pegándola con materiales del tipo cristal o metálicos Au-Sn.

14.2.4.2 Empaquetamiento plástico

Existen dos tipos de dispositivos encapsulados plásticos, los de post-molde como los de la figura y los de pre-molde como los de la figura . Los encapsulados de post-molde utilizan silicona termosellante, pegamento de silicona o resina de silicona y se moldean alrededor de la estructura principal del chip ensamblado después de que el chip se haya unido a la estructura principal. El proceso de post-molde es bastante agresivo. Para evitar la exposición del chip y uniones de los cables y tapa se desarrollo el método de pre-molde para empaquetar. El proceso de pre-molde el empaquetado es moldeado primero y después se emplaza el chip

y se hace sus interconexiones o cableado. El molde se realiza con material termosellante como los mencionados anteriormente o con polímeros termoplásticos. El empaquetado plástico de premolde es el equivalente al empaquetado de cavidad cerámica refractaria.

14.2.5 Otras técnicas de empaquetamiento

Las técnicas anteriores de empaquetamiento consumen mucho tiempo al unir los cables individualmente existen varias técnicas de empaquetamiento que son mas rápidas. Estas son:

1) El empaquetado TCP (tape carrier package)
2) El empaquetado de vuelta de chip (Flip Chip Pakage)

14.2.5.1 Empaquetado TCP

Esta técnica utiliza un conjunto de laminas finas de metal normalmente de cobre recubierto de oro o estaño en el empaquetado, en el en vez de cables. Estas patillas de metal se conectan al los contactos del chip, para ello los contactos del chip son recubierto por oro en forma de pequeño cilindros. En otros casos a las finas laminas de metal se les añade una soldadura para luego unirlas directamente con los contactos de Aluminio del chip, en la siguiente figura se muestran los distintos tipos de laminas utilizadas.

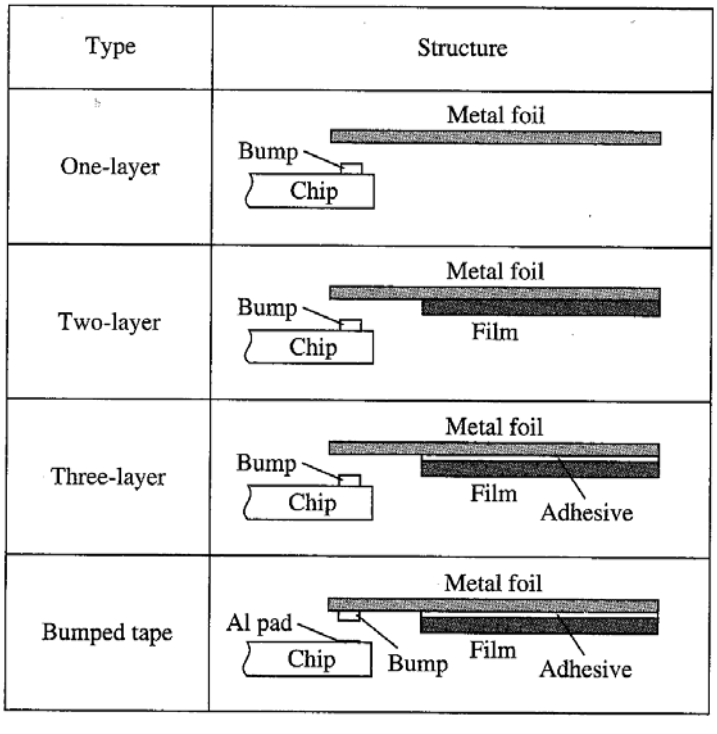


Figura 14.2.11: Diferentes métodos para realizar el empaquetado TCP.

Las uniones se hacen todas a la vez por termocompresión. Con esta técnica se pueden conseguir espaciados entre líneas inferiores a 100micras lo que lo hace adecuado para dispositivos con gran número de entradas salidas. Los problemas de esta técnica son el alto coste y la tecnología de empaquetamiento del siguiente nivel. El empaquetamiento usual en TCP es el plástico.

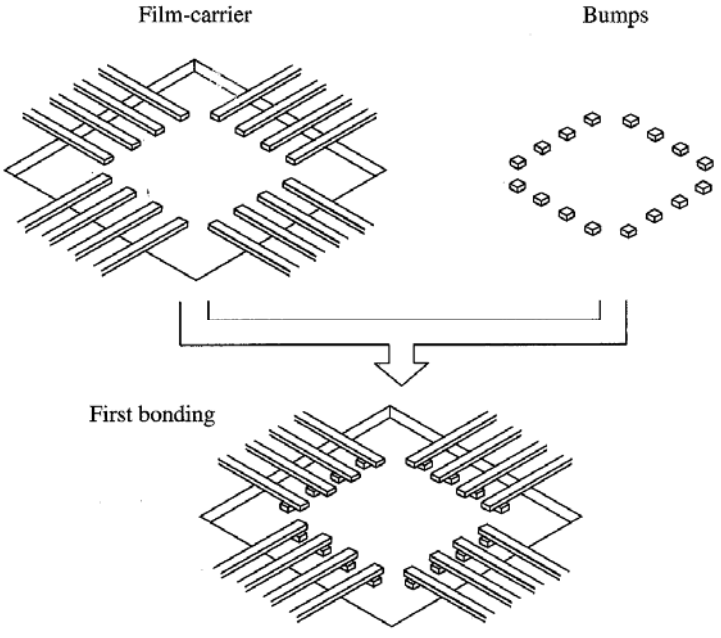


Figura 14.2.12: Esquema del procedimiento de empaquetado TCP.

14.2.5.2 Metodo flip-chip

La técnica flip-chip consiste en depositar metal (típicamente Pb o Sn) en forma de burbuja sobre los contactos de aluminio del chip antes de que estos se separen de la oblea. Estas soldaduras servirán para ser conectadas al substrato del empaquetado. Después de separar el chip de la oblea se pone boca abajo con las burbujas perfectamente alineadas con las metalizaciones del substrato. Después mediante unión ultrasónica se une cada burbuja del chip a su correspondiente conexión del substrato. La ventaja de este método es que todas las conexiones se realizan a la vez, se pude conseguir una gran densidad de interconexiones y se tiene una inductancia muy baja de interconexión con el empaquetado. Las desventajas se deben a que las uniones se hacer debajo del chip siendo imposible su inspección visual. Además de que tiene un peor comportamiento térmico.

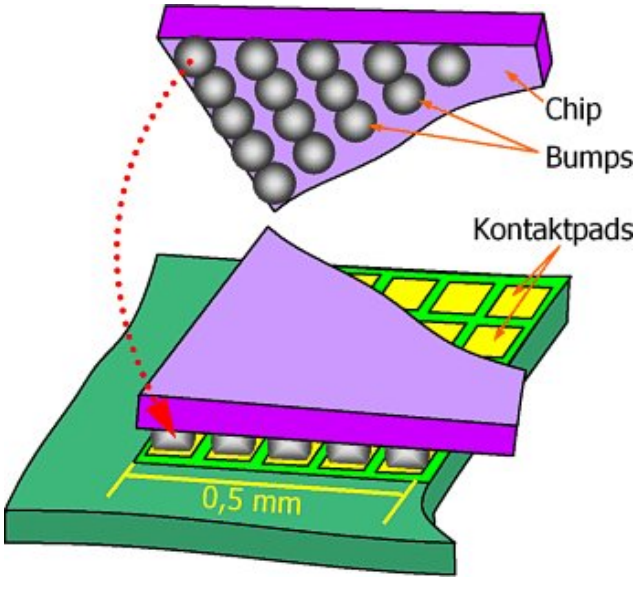


Figura 14.2.13: Esquema de la unión de los contactos del chip con los terminales del encapsulado mediante la técnica flip chip

www.mm.fh-heilbronn.de/wehl/ projekte/lotdruck.htm

http://education.netpack-europe.org/
http://extra.ivf.se/ngl/

.

14.3 Rendimiento y fiabilidad

Dos condiciones tienen que cumplir la tecnología electrónica la primera, la tecnología tiene que ser capaz de fabricar un circuito integrado en grandes cantidades y con un coste competitivo, la segunda el circuito debe de ser capaz de realizar su función durante toda la vida media estimada. Para producir circuitos que cumplan estas dos condicione se tendrán que conocer los mecanismos que hacen que los dispositivos sean caros y poco fiables.

El tamaño óptimo de un circuito integrado con respecto al número de funciones dependerá de varios factores: el número de chips del sistema, el coste esperado de un buen circuito, el coste de ensamblado y empaquetado y sobre todo la fiabilidad del sistema completo. Se tendrá que evaluar que es más rentable utilizar un número grande de CI pequeños o usar un menor número de CI de tamaño mayor.

Empezaremos viendo los mecanismos causantes de la perdida de rendimiento en la tecnología electrónica

14.3.1 Mecanismos de pérdida de rendimiento

Idealmente se esperaría que todos los chips que se producen en una misma oblea fueran buenos y funcionales. En la práctica la cantidad de chips en buen estado puede varia entre unos pocos o cantidades mas o manos cercanas al 100%. Las causas de chips en mal estado se pueden dividir en tres categorías: los problemas de procesado, los problemas de diseño y defectos aleatorios

14.3.1.1 Problemas de procesado

Si uno observa la fotografía o plano de una oblea testada se puede dar cuenta que esta tiene regiones con gran calidad de chips en buen estado mientras que existen otras donde casi todos los chips son defectuosos. Ver **Figura 14.3.1**.

La existencia de estas regiones de baja productividad se pueden deber a varias causas: variaciones en el espesor de las capas de oxido y polisilicio, en la resistencia de las capas implantadas, en la anchura de los rasgos definidos litográficamente, en la relación de una fotomascara con la anterior. Muchas de estas relaciones dependen unas de otras. Por ejemplo las regiones donde la capa de polisilicio es mas delgada que la media son grabadas de forma que las puertas de polisilicio serán mas pequeñas en esa región. Este efecto hace que haya una región donde las longitudes de canal sean muy pequeñas y los dispositivos no dejen de conducir cuando se les aplique la apropiada tensión de puerta. Por tanto estos dispositivos tendrán excesivas corriente de perdidas o no funcionaran. Las variaciones en el dopado de las capas implantadas pueden provocar variaciones en la resistencia de contacto de la capa implantada. Variaciones del espesor de un material dieléctrico puede

provocar variaciones en el tamaño de las ventanas de contacto. Ambos efectos puede conducir a circuitos no operativos su los caminos incluidos en el circuito no tienen un bajo valor de resistencia de contacto.

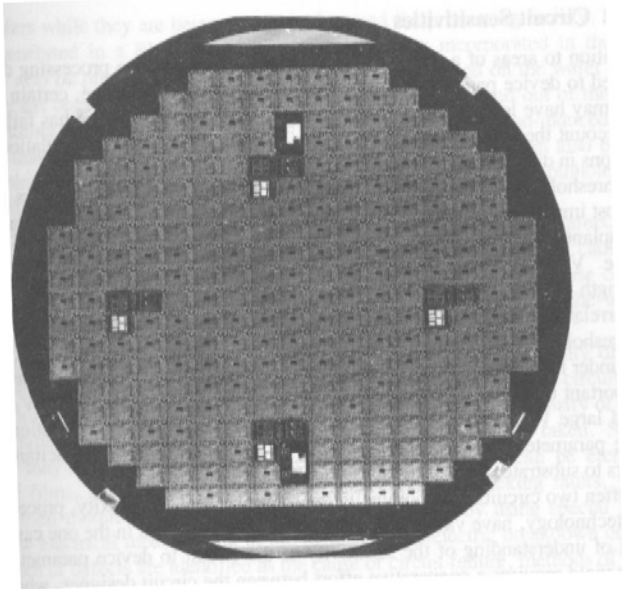


Figura 14.3.1 fotografía de una oblea con chips defectuosos señalados en oscuro.

Durante la fabricación de la oblea se llevan a cabo varias operaciones que pueden dar lugar a pequeñas pero críticas variaciones en el espesor y tamaño de la oblea. Por ejemplo durante el proceso de oxidación térmica se crece una capa de oxido por las dos caras de la oblea que tensionará la capa de silicio de forma que la oblea resultante de la oxidación tendrá un diámetro mayor que el anterior. Si el grado de tensión supera el límite elástico del material se producirá una deformación. Si el oxido es quitado de unas de las caras la oblea estará curvada.

A medida que las técnicas de procesado y procesos se van desarrollando muchas de estos efectos se reducen o son eliminados pero sin embargo otros nuevos aparecen.

14.3.1.2 Problemas de diseño

Además de las zonas de la oblea donde el rendimiento es bajo debido a las dificultades del proceso que hacen que los parámetros de los dispositivos se salgan del rango de especificaciones, algunas áreas de la oblea pueden tener un bajo rendimiento debido a que el diseño del circuito falla al no tener en cuenta la variación esperada de los parámetros del dispositivo y la relación entre las variaciones de los diferentes parámetros.

En el caso de transistores MOSFET la tensión umbral (V_T) y la longitud de canal (L) son dos de los parámetros más importantes del

dispositivo en el diseño del circuito integrado. Las variaciones en el dopado de substrato, en la dosis de dopado y el espesor del oxido de puerta causaran variaciones en la tensión umbral. Las variaciones en la longitud de la puerta y las profundidades de las difusiones de fuente y drenador hacen que la longitud del canal varié. Las variaciones de la tensión umbral y de la longitud de canal no están generalmente relacionadas una con otras. Sin embargo la velocidad de un circuito integrado normalmente se incrementa a medida que decrece la tensión umbral y la longitud de puerta. El funcionamiento del circuito se simula a menudo bajo condiciones de alta velocidad (V_T y L pequeña) y baja velocidad (V_T y L grande). Es importante que la operación del circuito se simule en el caso de V_T pequeño y L grande y en el caso de V_T grande y L pequeño. El diseño del circuito debe también considerar las variaciones de otros parámetros del circuito como la resistencia de las regiones implantadas, las capacidades de los conductores con el substrato, las resistencias de contacto y las corrientes de perdidas.

A menudo dos circuitos con el mismo tamaño nominal y complejidad, procesados bajo la misma tecnología tienen rendimientos de producción muy diferentes. La baja productividad en unos casos se debe al desconocimiento de la sensibilidad de los parámetros del circuito. Un mayor rendimiento de la productividad requiere de la cooperación entre el diseñador de circuitos, quien identifica los parámetros del dispositivo para los cuales el circuito es sensible y el ingeniero de proceso, el cual optimiza el valor y el rango de variación de esos parámetros. Una vez que se han determinado las sensibilidades del circuito para los parámetros específicos del proceso el rediseño del circuito para reducir esas sensibilidades conseguirá un alto rendimiento en la producción y un bajo coste con una atención mínima del ingeniero de proceso.

14.3.1.3 Defectos puntuales

Dentro de las áreas de la oblea que están correctamente fabricadas, todos los paramentos de proceso están dentro del rango de operación correcta de los circuitos se pueden encontrar algún chip en mal estado. La causa de esta perdida de rendimiento de fabricación se debe a los defectos puntuales.

Un defecto de punto es una región defectuosa de la oblea que tiene un tamaño inferior al de un chip. Por ejemplo consideremos un chip de 2000μm cuadradas con una pista de 2 micras. Una partícula de polvo de tres micras de diámetro sobre la oblea puede causar la rotura de el metal conductor.

Existen varios tipos de defectos de proceso que se pueden considerar como defectos puntuales. Uno de los mas usuales son la partículas de polvo u otro tipos de partículas del ambiente. Estas partículas pueden caer sobre la oblea mientras son trasportadas por las instalaciones, o pueden ser generados en la operación de deposición de

película delgada. Estas partículas también pueden estar presentes en las películas fotoresistivas y ser depositadas durante el proceso de litografía o pueden ser partículas de silicio desprendidas de la oblea durante su manejo que quedan adheridas a la superficie de la oblea.

Los defectos de punto pueden deberse a las partículas que se adhieren a las mascaras litográficas durante la generación de estas. Lo que produce un error permanente en la mascara. O partículas que se adhieren a estas mascarás durante el proceso litográfico. Para evitar esto las mascaras tienen que ser limpiadas periódicamente.

Un fabricación satisfactoria requiere el continuo control de la densidad de defectos puntuales. Este control se realiza mediante la observación con microscopio SEM durante todos los pasos del proceso de fabricación.

14.3.2 Fiabilidad

La fiabilidad de los circuitos integrados viene determinada por varias disciplinas como el diseño, los procesos de fabricación del chip, el empaquetamiento y testeo del C.I. Se usan procesos de fabricación y manufacturado delicados para producir finalmente un CI. Cuando este falla se puede deber a una gran cantidad de causas. Se requieren sofisticados y tediosos análisis para saber la causa del error. Desde la perspectiva del consumidor lo que el requiere es que sea útil durante la vida media especificada. Normalmente los fabricantes diseñan un CI para 10 años. La fiabilidad del producto esta asegurada si cada uno de los elementos del producto es fiable durante esos años. Existen tres elementos fundamentales en la fiabilidad: la fiabilidad del diseño, la fiabilidad de los procesos, y la fiabilidad del ensamblado. Si la fiabilidad de cada uno de esos procesos cumple los requisitos de vida media, entonces la fiabilidad del producto esta asegurada.

La tecnología ULSI conlleva el escalado de los procesos por debajo de las dimensiones submicra, como también la adición de nuevos módulos de procesos no usados en la era VLSI. La fiabilidad de cada uno de esos nuevos procesos y como interacciona con los anteriores será critica en la fiabilidad final del proceso completo. En la tecnología ULSI el concepto de diseño para la fiabilidad es muy importante. Se debe hacer un diseño de la fiabilidad durante todos los pasos de fabricación de un CI, diseño del circuito, procesado y manufacturado.

Los circuitos integrados están formados por varios elementos discretos como transistores, resistencias interconexiones, películas de dieléctricos y capacidades. Todas estas partes de un CI integrado sufren poco a poco un deterioro debido a su operación. En el caso de los transistores una de las causas de fallo es el efecto de los electrones calientes “Hot carriers”, otra, la ruptura de oxido de puerta. En el caso de las interconexiones aparecen los problemas de electromigración y migración de estrés.

REFERENCIAS

[1] "VLSI Technology", Sze, Ed. Mcgraw-Hill

[2] "ULSI Technology", Chang and Sze, Ed. Mcgraw-Hill

[3] "Solid State Electronic Devices" Streetman and Banerjee, Prentice Hall, 2000 Fifth edition

[4] Glosario: http://semiconductorglossary.com/

[5] Microscopia óptica:
http://www.microscopy.fsu.edu/primer/techniques/dic/dicintro.html

[6] Microscopia electrónica Universidad de Nebraska:
http://www.unl.edu/CMRAcfem/

[7] IBM Análisis de materiales:
http://www.almaden.ibm.com/st/scientific_services/materials_analysis/

[8] Centro de Micro-Análisis de Materiales CMAM:
http://www.uam.es/otroscentros/cmam/espanol/tecnicas/node1.html

[9] Preparación de obleas, IC Services Corporation:
http://www.icservice.com/

[10] The Nordic Electronics Packaging Guideline: http://extra.ivf.se/ngl/

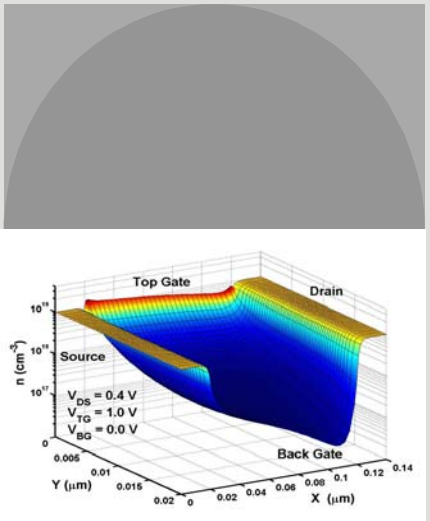
[11] Netpack: http://education.netpack-europe.org/

[12] Video de la empresa Surftape, (requiere Quicktime):
http://www.surftape.com/solutions2.html

15

Capítulo

Modelos para Simulación de Dispositivos Electrónicos



Concentración de electrones en un DGSOI

Índice

Objetivos

- Justificar la necesidad de simuladores de dispositivos
- Presentar las diferentes aproximaciones a la ecuación de transporte
- Introducir el método Monte Carlo
- Presentar diferentes modelos de Movilidad y Generación-Recombinación
- Estudiar el modelo de discretización del espacio continuo y de las ecuaciones a resolver
- Justificar los diferentes tipos de condiciones de contorno que pueden ser impuestas al espacio a simular
- Presentar diferentes soluciones comerciales

Palabras Clave

Ecuación de Poisson
Ecuación de Continuidad
Ecuación de Transporte de Boltzmann
Modelo de Difusión-Deriva
Relación de Einstein
Modelo Hidrodinámico
Método de Monte Carlo
Procesos de dispersión
Autoscattering
Aproximación semiclásica
Hipótesis ergódica
Ecuación de Schrödinger
Modelo de Caughey-Thomas
Modelo Shockley-Reed-Hall
Recombinación Auger
Método de Diferencias Finitas
Condición de contorno Dirichlet
Condición de contorno Neumann

15.1 Modelado de Dispositivos

Modelado de Dispositivos
Toda serie de modelos y métodos que describen el transporte de portadores, distribuciones de potencial y de campos en una estructura a estudio.

La simulación se ha convertido en un campo muy importante para el estudio de dispositivos semiconductores. La complejidad cada vez mayor de las estructuras, la miniaturización de las dimensiones y la utilización de efectos más complejos para mejorar las prestaciones hacen necesario un tratamiento mucho más riguroso de los modelos que rigen el comportamiento de los dispositivos. Las aproximaciones realizadas para los estudios teóricos dejan de tener validez y el problema a resolver deja de tener solución analítica haciéndose necesario el uso de métodos numéricos que permitan alcanzar una solución satisfactoria al problema.

A pesar de todo, existen diferentes niveles de aproximación al problema debido a que, para la resolución, se necesita una gran cantidad de recursos en cuanto a potencia de cálculo y tiempo se refiere requiriéndose un compromiso entre exactitud y precisión en la solución y tiempo de simulación. Los modelos que se pueden encontrar en las distintas herramientas van desde los de más simples de difusión y deriva a los más complejos y costosos en cuanto a requerimientos de cálculo como pueden ser los de balance de energía para la resolución de la ecuación de transporte de Boltzmann (BTE). Asimismo la complejidad de la física puesta en juego hace necesario el uso de códigos de tipo Monte Carlo que resuelven de una manera estocástica la BTE y la resolución de la ecuación de Schrödinger para tener en cuenta diferentes efectos cuánticos que cada día son más importantes para explicar el comportamiento de dispositivos con dimensiones submicra o nanométricas.

Las herramientas de simulación son ampliamente utilizadas en estudios de escalado de dispositivos y optimización de tecnologías tanto existentes como emergentes. Por tanto, la capacidad de estos programas de representar las prestaciones actuales y de predecir las de futuras tecnologías y sus limitaciones es de vital importancia ya que permite a las compañías ahorrar grandes cantidades de dinero en los procesos de desarrollo antes de la fabricación en masa de los distintos componentes y a los centros de investigación comprobar la viabilidad teórica de dispositivos basados en efectos físicos novedosos con geometrías diferentes a las configuraciones estándar.

El tipo de simulador elegido para cada caso dependerá pues del problema a tratar, de la precisión que se quiera conseguir en los cálculos y de los efectos que se quieran tener en cuenta. Al mismo tiempo existe una limitación a la hora de elegir el método en función de los medios computacionales disponibles y el tiempo que pueda dedicarse al estudio. Todos estos factores deben llevar a una solución de compromiso a decidir por el usuario.

15.2 Clasificación de los Simuladores

Con todas estas consideraciones es fácil de imaginar que no es posible ni práctico realizar una clasificación única de los distintos tipos de simuladores. A continuación se presentan algunas posibles clasificaciones atendiendo a diferentes aspectos como pueden ser la geometría a simular o el modelo utilizado para la descripción de los fenómenos de transporte.

En primer lugar uno debe considerar qué tipo de estructura va a ser simulada, de forma que se pueda decidir qué formulación de las ecuaciones puestas en juego es la más interesante pudiéndose elegir entre modelos uni, bi y tridimensionales. Normalmente se intenta usar la descripción que tenga en cuenta los efectos que quieren ser simulados pero sin sobredimensionar el problema para no obtener un código excesivamente costoso en tiempo de computación. Así por ejemplo, una simulación monodimensional puede ser suficiente para estudiar una unión MOS en equilibrio mientras que para el caso de un transistor MOSFET con tensión aplicada entre drenador y fuente será necesaria una simulación, al menos, bidimensional para el caso en que la longitud de canal sea pequeña.

Otra decisión crítica que debe tomarse es la elección del modelo que describirá el comportamiento del sistema. En la mayoría de los casos la distribución del potencial en el dispositivo viene descrita por la ecuación de Poisson en cualquiera de sus versiones

$$\nabla\cdot\left(\varepsilon\nabla V\right)=-\rho. \tag{15.1}$$

Sin embargo existe mucha más flexibilidad a la hora de elegir el modelo que describirá el transporte de los portadores de carga. La teoría semi-cásica de transporte está basada en la ecuación de transporte de Boltzmann (BTE)

$$\begin{aligned}&\frac{\partial f}{\partial t}+\boldsymbol{v}\cdot\nabla_{\boldsymbol{r}} f+\frac{e\boldsymbol{E}}{\hbar}\cdot\nabla_{\boldsymbol{k}} f=\\&=\sum_{\boldsymbol{k}'} S\left(\boldsymbol{k}',\boldsymbol{k}\right) f\left(\boldsymbol{r},\boldsymbol{k}',t\right)\left[1-f\left(\boldsymbol{r},\boldsymbol{k},t\right)\right]-\\&-\sum_{\boldsymbol{k}'} S\left(\boldsymbol{k},\boldsymbol{k}'\right) f\left(\boldsymbol{r},\boldsymbol{k},t\right)\left[1-f\left(\boldsymbol{r},\boldsymbol{k}',t\right)\right]\end{aligned} \tag{15.2}$$

donde **r** representa la posición, **k** el momento, $f\left(\boldsymbol{k}',t\right)$ la función de distribución (por ejemplo la de Fermi-Dirac si se está en equilibrio), $\boldsymbol{v}$ la velocidad de grupo, **E** el campo eléctrico, $S\left(\boldsymbol{k},\boldsymbol{k}'\right)$ la probabilidad de transición entre dos estados con momento **k** y **k'** y $1-f\left(\boldsymbol{k}',t\right)$ la probabilidad de no ocupación del estado con momento **k'**. La sumatoria del término derecho de la ecuación representa el término de colisiones que tiene en cuenta los diferentes procesos de dispersión. Los del lado

izquierdo expresan las dependencias de la función de distribución con el tiempo, el espacio y el momento.

La BTE es válida asumiendo transporte semiclásico y aproximación de masa efectiva (en la que se incorporan los efectos cuánticos de la periodicidad del cristal).

Es posible obtener soluciones analíticas de esta ecuación solo bajo condiciones muy restrictivas por lo que, en la práctica, se suelen utilizar diferentes aproximaciones que intentan evitar su resolución directa.

El Modelo de Difusión-Deriva para el Transporte

Una de las más utilizadas por su relativa sencillez es la ecuación de continuidad en su aproximación de difusión-deriva que puede ser expresada como:

$$\frac{\partial n}{\partial t} = \frac{1}{q}\nabla \cdot \boldsymbol{J_n} + U_n \tag{15.3}$$

$$\frac{\partial p}{\partial t} = -\frac{1}{q}\nabla \cdot \boldsymbol{J_p} + U_p \tag{15.4}$$

donde n y p representan la concentración de electrones y huecos respectivamente, U las tasas de Generación-Recombinación de portadores y $\boldsymbol{J}$ las densidades de corriente cuya expresión viene dada por:

$$\boldsymbol{J_n} = qn\mu_n \boldsymbol{E} + qD_n \nabla n \tag{15.5}$$

$$\boldsymbol{J_n} = qp\mu_p \boldsymbol{E} - qD_p \nabla p \tag{15.6}$$

El primer término de la densidad de corriente corresponde a la corriente de arrastre y depende del campo eléctrico y de la movilidad de los portadores. El segundo, describe los procesos de difusión originados por la diferencia en la concentración de portadores en las diferentes zonas del dispositivo.

Si se recuerda la definición de campo eléctrico

$$\boldsymbol{E} = -\nabla V \tag{15.7}$$

se encuentra que el cálculo de las concentraciones de portadores depende también de la distribución de potencial con lo que aparece un sistema de ecuaciones diferenciales acoplado, requiriéndose una resolución autoconsistente junto a la ecuación de Poisson.

El Modelo Hidrodinámico

Una segunda aproximación consiste en el modelo hidrodinámico que expresa la conservación de partículas, de momento y energía a partir del cálculo de los momentos de la BTE en términos de la función de distribución para los portadores que vienen definidos como

$$n = \int f(\mathbf{r},\mathbf{k},t)\, d^3\mathbf{k} \tag{15.8}$$

$$n\mathbf{v} = \int \mathbf{v}(\mathbf{k})\, f(\mathbf{r},\mathbf{k},t)\, d^3\mathbf{k} \tag{15.9}$$

$$W = \frac{m^*}{2} \int \mathbf{v}^2(\mathbf{k})\, f(\mathbf{r},\mathbf{k},t)\, d^3\mathbf{k} \tag{15.10}$$

donde n es la concentración de partículas, $\mathbf{v}$ es la velocidad media y W la densidad de energía cinética.

Este modelo es bastante más complicado que el comentado anteriormente de difusión y deriva. Uno de los principales problemas de este método es que las tres ecuaciones básicas de conservación pueden ser resueltas siempre que uno conozca con un grado de exactitud adecuado una aproximación de las funciones de distribución de forma que los momentos puedan ser calculados de manera razonable. Incluso teniendo esta información, la tarea de resolver estas ecuaciones es muy costosa computacionalmente, sobre todo para aplicaciones en ingeniería.

El Método Monte Carlo

Quizás el método más utilizado para el estudio del transporte en semiconductores junto al de difusión y deriva sea el Monte Carlo. Bajo este término se agrupan una serie de técnicas de naturaleza estocástica que utilizan la generación de números aleatorios. Para el caso del transporte se usa para resolver la BTE sin realizar ninguna suposición sobre la función de distribución. Las distribuciones de probabilidad apropiadas para los distintos parámetros se obtienen a partir de la generación de secuencias números aleatorios uniformemente distribuidos.

Las aproximaciones tomadas para el modelo semiclásico son las siguientes:

- Las partículas se consideran como objetos puntuales con posición y momentos perfectamente definidos violándose por tanto, el principio de incertidumbre. Esta suposición es descartada en el caso de la interpretación cuántica.
- Se hace uso de la aproximación de masa efectiva, donde se incluyen los efectos cuánticos derivados de la existencia de un potencial periódico correspondiente a la estructura cristalina
- El movimiento de las partículas se describe como una serie de vuelos libres interrumpidos por procesos de dispersión. Las variaciones en la trayectoria y en el momento se calculan utilizando mecánica clásica, mientras que las probabilidades de dispersión se calculan de acuerdo a las reglas de la mecánica cuántica.

Dentro de estas técnicas se pueden encontrar dos aproximaciones principales para la descripción del sistema.

La primera denominada *Monte Carlo de una Partícula* se utiliza para el estudio de las propiedades de los sistemas en estado estacionario. Suponiendo que se conoce la distribución de potencial para ese estado, se estudia el movimiento de una partícula durante un tiempo lo suficientemente elevado. De esta forma, aplicando la hipótesis ergódica, se pueden relacionar las magnitudes medias temporales de la partícula con los valores medios de esas magnitudes calculados para toda la colectividad de partículas en un momento dado se cumple que

$$\overline{f} = \langle f \rangle \tag{15.11}$$

donde f representa la magnitud a estudio. Este método tiene una aplicación limitada en el estudio de dispositivos, aunque permite obtener resultados interesantes para dependencias de la movilidad y otras magnitudes con respecto al campo eléctrico aplicado de forma que pueden obtenerse modelos para su posterior uso en otros simuladores basados en la BTE como es el caso de los de difusión y deriva.

La segunda aproximación se conoce como *Ensemble Monte Carlo*. El algoritmo utilizado es prácticamente el mismo que para el caso anterior con la salvedad de que se simula a la vez un conjunto grande de partículas que representa a la totalidad de partículas. Este método es susceptible de ser paralelizado para poder aprovechar las posibilidades de la supercomputación. Debido a que se realiza un estudio de muchas partículas es posible calcular valores promedio directamente para cada intervalo temporal, con lo que no es necesaria la condición de estado estacionario para poder aplicar este método. Por todos estos motivos, es el mejor candidato para la simulación de procesos transitorios.

Una variante de este método denominada autoconsistente resulta ser la más adecuada para el estudio de dispositivos. En ella la ecuación de Poisson se vuelve a resolver cada cierto intervalo de tiempo (típicamente del orden de los femtosegundos) de forma que se tienen en cuenta las variaciones que en la distribución de potencial provoca el movimiento de los portadores dentro del dispositivo. Cuando las características del dispositivo a estudio lo requieren, puede que sea necesario unir a la solución autoconsistente la ecuación de Schrödinger, con lo que las distribuciones de portadores y energías antes de iniciar EMC en cada momento vienen dadas a partir de las funciones de onda obtenidas de resolver dicha ecuación.

A continuación se presenta una breve descripción del método con el que se deciden la duración de los vuelos libres entre colisiones, el mecanismo de scattering que actúa y el estado final tras la dispersión.

Para el cálculo del tiempo de vuelo libre se debe tener en cuenta una serie de consideraciones. En primer lugar el vector de onda **k** cambia continuamente debido al campo eléctrico aplicado. Si la probabilidad de que un electrón con vector de onda **k** sufra una dispersión en dt es

$\Gamma\left(\boldsymbol{k}\left(t\right)\right)dt$, la probabilidad de que, habiendo sufrido una colisión en t=0, sufra una en t es:

$$\exp\left\{-\int_0^t \Gamma\left(\boldsymbol{k}\left(t'\right)\right)dt'\right\} \tag{15.12}$$

y por tanto la probabilidad de que un electrón sufra una colisión en un intervalo dt alrededor de un instante de tiempo t viene dada por

$$P\left(t\right)dt = \Gamma\left(\boldsymbol{k}\left(t\right)\right)\exp\left\{-\int_0^t \Gamma\left(\boldsymbol{k}\left(t'\right)\right)dt'\right\}dt \tag{15.13}$$

En general la forma de $\Gamma\left(\boldsymbol{k}\right)$ es muy complicada, de modo que la integral resulta bastante difícil de resolver. Sin embargo en 1968 Rees propuso la solución mediante la introducción de un proceso de dispersión ficticio denominado "autoscattering". Se toma $\Gamma_M \equiv \frac{1}{\tau_0}$ definido como el valor máximo de $\Gamma\left(\boldsymbol{k}\right)$ en la región de interés. De esta forma se tiene que para cualquier instante, la probabilidad total de dispersión es siempre constante e igual a Γ_M. Cuando el electrón sufre un autoscattering su estado después de la colisión es el mismo que el inicial con lo que sigue su movimiento como si nada hubiese pasado. Aplicando este razonamiento la ecuación (15.13) queda de la forma

$$P\left(t\right)dt = \frac{1}{\tau_0}e^{-\frac{t}{\tau_0}} \tag{15.14}$$

y la duración del vuelo libre

$$t_r = -\tau_0 \ln\left(r\right) \tag{15.15}$$

donde *r* es un número aleatorio uniformemente distribuido.

Una vez que se ha determinado el tiempo de vuelo se debe determinar qué mecanismo de dispersión entra en juego para poder calcular así el estado final tras la colisión. Debido a que la probabilidad de que un vuelo termine como consecuencia del mecanismo n-ésimo es proporcional a $\Gamma_n\left(\boldsymbol{k}\right)$ y

$$\Gamma_M = \sum_n \Gamma_n\left(\boldsymbol{k}\right) \tag{15.16}$$

el mecanismo responsable del fin del vuelo libre puede determinarse eligiendo un número aleatorio uniformemente distribuido entre 0 y Γ_M de forma que el mecanismo seleccionado es aquel para el que el número aleatorio *r* es menor que la primera suma parcial $\sum_j \Gamma_j$ que se encuentre:

$$\sum_{j=1}^{N-1}\Gamma_j(\boldsymbol{k}) < r < \sum_{j=1}^{N}\Gamma_j(\boldsymbol{k}). \qquad (15.17)$$

Una vez seleccionado el mecanismo de dispersión el estado final es calculado en función del modelo que se esté utilizando para dicho proceso.

15.3 Modelos de Movilidad y Generación-Recombinación

Debido a su mayor sencillez y a su velocidad de resolución, el modelo de transporte más usado (aunque no el único) en la industria para simular características de dispositivos es el de Difusión y Deriva. Como se puede comprobar en las ecuaciones (15.3) - (15.6) existen tres términos sobre los que no se ha comentado nada y que resultan parámetros fundamentales en la resolución del sistema. Estos son la movilidad, μ, el coeficiente de difusión, D, y la tasa de Generación-Recombinación, U.

El coeficiente de difusión se suele relacionar con la movilidad a través de la relación de Einstein que es válida sólo para en condiciones de equilibrio o bajos campos:

$$\frac{D}{\mu} = \frac{kT}{q}. \qquad (15.18)$$

Para el caso de la movilidad y los términos de G-R existen diferentes modelos que permiten ampliar el rango de validez de la aproximación de Difusión Deriva.

Modelos de Movilidad

La movilidad es una magnitud que da idea de la facilidad con que se pueden mover los portadores en un cristal semiconductor. Su valor no es constante en todo el dispositivo y depende de múltiples factores relacionados tanto con las características del dispositivo como con la polarización aplicada en los terminales.

En el caso de transporte en volumen la movilidad viene determinada principalmente por la dispersión producida por los átomos de la red y las impurezas ionizadas con valores típicos del orden de 1350 cm^2/Vs para electrones y de 500 cm^2/Vs para huecos para un cristal ligeramente dopado a 300 K Para el caso de láminas de inversión en un MOSFET (gases de e-h bidimensionales) predominan los efectos producidos por los campos, tanto longitudinal como transversal, y la rugosidad superficial. La movilidad disminuye a valores del orden de 670 cm^2/Vs para electrones y de 160 cm^2/Vs para huecos para $V_{GS}=V_T$ y N_A , $N_D < 10^{17}\ cm^{-3}$.

En dispositivos en los que aparecen láminas de inversión, los portadores que circulan por ella están sometidos a campos transversales muy fuertes producidos por la polarización de la puerta. La dependencia

con el campo transversal puede modelarse de diferentes maneras. Por ejemplo introduciendo modelos semiempíricos que permiten obtener un valor de movilidad efectiva que será luego corregida por los efectos de campo longitudinal.

Los modelos semiempíricos se basan en el concepto de campo efectivo que se define como el campo medio en la lámina de inversión. Uno de los más usados es el que fue propuesto por Liang et al en 1986 cuya expresión es:

$$\mu_{eff} = \frac{\mu_0}{1 + \left(\frac{E_{eff}}{E_{crit}} \right)^{\nu}} \tag{15.19}$$

donde μ_0 representa la movilidad a bajo campo y E_{crit} y ν son parámetros obtenidos de forma empírica.

Un segundo modelo, esta vez obtenido de forma empírica, es el de Lombardi que incorpora efectos tales como son los de dispersión superficial por fonones acústicos ($\mu_{ac,}$) rugosidad superficial ($\mu_{sr,}$) y efectos de transporte en volumen ($\mu_{b,}$). En este caso las dependencias con el campo transversal son directas y no a través del campo medio. La movilidad total puede ser calculada usando la regla de Mathiessen

$$\mu_{eff} = \left[\frac{1}{\mu_{ac}} + \frac{1}{\mu_b} + \frac{1}{\mu_{sr}} \right]^{-1} \tag{15.20}$$

Una vez obtenida la movilidad efectiva por cualquiera de los modelos que se elija, se introduce ésta en la expresión que contiene la dependencia longitudinal del campo. El modelo más utilizado es el de Caughey-Thomas

$$\mu = \frac{\mu_{eff}}{\left[1 + \left(\frac{\mu_{eff} E_{\square}}{v^{sat}} \right)^{\beta} \right]^{\beta}} \tag{15.21}$$

donde v^{sat} es la velocidad de saturación para los portadores y β un parámetro empírico (típicamente $\beta = 2$ para electrones y $\beta < 2$ para huecos).

Modelos de Generación-Recombinación

En cuanto a los términos de generación-recombinación, existen diferentes modelos según el mecanismo que entre en juego para describir el proceso. Así, el modelo de Shockley-Reed-Hall se utiliza cuando los procesos de G-R están relacionados con niveles creados por trampas electrónicas.

$$U_{SRH} = \frac{np - n_i^2}{\tau_p \left[n + n_i \, e^{\left(\frac{q(E_t - E_i)}{kT} \right)} \right] + \tau_n \left[p + p_i \, e^{\left(\frac{q(E_i - E_t)}{kT} \right)} \right]} \qquad (15.22)$$

donde τ_n y τ_p son los tiempos de vida media para electrones y huecos respectivamente y E_t representa el nivel de energía de la trampa envuelta en el proceso.

Para el caso de zonas con un dopado muy alto como puede ser la de emisor en un transistor bipolar se utiliza la denominada recombinación Auger

$$U_{Aug} = C_n \left[pn^2 - nn_i^2 \right] + C_n \left[np^2 - pn_i^2 \right]. \qquad (15.23)$$

Finalmente existe un tercer modelo que da cuenta de los efectos de la ionización por impacto cuya dependencia con los campos viene dada por:

$$U_I = \frac{a_n^{\infty} \exp\left(-\frac{E_n^{crit}}{E} \right)^{\beta_n} \left| \boldsymbol{J_n} \right| + a_p^{\infty} \exp\left(-\frac{E_p^{crit}}{E} \right)^{\beta_p} \left| \boldsymbol{J_p} \right|}{q} \qquad (15.24)$$

15.4 Discretización del Espacio Continuo

Todas las expresiones anteriormente propuestas están formuladas en un espacio continuo, sin embargo, a la hora de ser resueltas numéricamente tanto las variables como los operadores que entran en juego deben ser transformados a una formulación discreta debido a la incapacidad de los computadores de realizar cálculos en el continuo.

El problema de la discretización no resulta trivial y de él depende en buena medida que con la simulación se obtengan unos resultados aceptables o que no resulte prohibitiva en términos de tiempo de simulación.

Existen distintas aproximaciones para la discretización del espacio. La primera de ellas, que será comentada con más detalle posteriormente, es la denominada de diferencias finitas, en la que el espacio es sustituido por un mallado de puntos.

Otra aproximación muy utilizada sobre todo en simuladores comerciales es la de elementos finitos. En ella el espacio es dividido en polígonos (normalmente triángulos y rectángulos) de forma que la región a estudio queda completamente teselada por ellos.

Por último existe un tercer método denominado espectral que se basa en algoritmos de resolución de ecuaciones a través de la transformada de Fourier.

En general se pueden definir tres conceptos que nos indican si el esquema de discretización es apropiado o no para el problema que se quiere resolver:

- Estabilidad
- Convergencia
- Consistencia

Estabilidad:
Se dice que un esquema es estable si la solución permanece acotada durante el proceso de resolución.
Convergencia:
Un esquema es convergente cuando la solución numérica se acerca a la solución real cuando el tamaño del mallado y el paso temporal tienden a cero.
Consistencia:
Un esquema es consistente si el error de truncamiento tiende a cero cuando el tamaño del mallado y el paso temporal tienden a cero.

El Método de Diferencias Finitas

El método de diferencias finitas es el más utilizado de todos los esquemas de discretización debido a su relativa simplicidad. Está basado en el uso de series de Taylor truncadas para realizar aproximaciones numéricas a las derivadas. Así

$$u(x+\Delta x) = u(x) + \Delta x \frac{\partial u}{\partial x} + \frac{\Delta x^2}{2}\frac{\partial^2 u}{\partial x^2} + O\left(\Delta x^3\right) \tag{15.25}$$

$$u(x-\Delta x) = u(x) - \Delta x \frac{\partial u}{\partial x} + \frac{\Delta x^2}{2}\frac{\partial^2 u}{\partial x^2} + O\left(\Delta x^3\right). \tag{15.26}$$

A partir de estas dos ecuaciones se puede obtener diferentes aproximaciones a la primera derivada

De (15.25) se tiene la aproximación de diferencias hacia delante:

$$\frac{\partial u}{\partial x} = \frac{u(x+\Delta x) - u(x)}{\Delta x} + O(\Delta x). \tag{15.27}$$

De (15.26) se tiene la aproximación de diferencias hacia atrás:

$$\frac{\partial u}{\partial x} = \frac{u(x) - u(x-\Delta x)}{\Delta x} + O(\Delta x). \tag{15.28}$$

Restando (15.25) - (15.26) se tiene la aproximación en diferencias centradas:

$$\frac{\partial u}{\partial x} = \frac{u(x+\Delta x) - u(x-\Delta x)}{2\Delta x} + O\left(\Delta x^2\right) \tag{15.29}$$

Una aproximación de la derivada segunda se puede obtener a partir de (15.25) + (15.26)

$$\frac{\partial^2 u}{\partial x^2} = \frac{u(x+\Delta x) - 2u(x) + u(x-\Delta x)}{\Delta x^2} + O\left(\Delta x^2\right). \tag{15.30}$$

Los términos en $O(\Delta x)$ y $O\left(\Delta x^2\right)$ indican el error de truncamiento que se comete en la aproximación. Cuanto mayor sea el orden del resto menor será el error de truncamiento. Estas aproximaciones son las más utilizadas aunque es posible obtener expresiones en las que se tengan en cuenta más puntos.

15.5 Condiciones de Contorno

La resolución de las ecuaciones para un dispositivo implica resolver un problema de condiciones de contorno. Estas dependerán del tipo de frontera ante la que nos encontremos y de la variable que se deba caracterizar. Las condiciones de contorno fijan los valores de las variables o de sus derivadas de forma que se cumplan ciertas leyes físicas o condiciones impuestas desde el esterior como pueden ser potenciales aplicados en los contactos.

En general se suelen distinguir los siguientes tipos de fronteras sobre las que se aplican condiciones de contorno:

- **Contactos Ohmicos:**

  Dentro de estos contactos se pueden encontrar dos tipos; controlados por tensión o por corriente. En el primero de ellos la condición para el potencial electrostático es de tipo Dirichlet y se expresa como

$$\psi(t) - \psi_b - \psi_D(t) = 0 \tag{15.31}$$

  donde ψ_b es el potencial debido al dopado de la zona de contacto y $\psi_D(t)$ el potencial externo aplicado.

  Para el caso de contactos controlados por corriente se tiene

$$\int_{\partial D_o} \left(\boldsymbol{J_n} + \boldsymbol{J_p} \right) \cdot \boldsymbol{dA} - I_D(t) = 0 \tag{15.32}$$

  Donde $I_D(t)$ es la corriente que se fuerza a pasar por el contacto.

  Para determinar la condición de contorno de los portadores se asume la condición de equilibrio térmico con lo que se cumplen las relaciones:

$$np - n_i^2 = 0 \tag{15.33}$$

$$n - p - C = 0 \tag{15.34}$$

$$C = N_D^+ - N_A^-. \tag{15.35}$$

  Las condiciones de contorno para portadores quedan de la siguiente forma

$$n = \frac{\sqrt{C^2 + 4n_i^2} + C}{2}. \tag{15.36}$$

$$p = \frac{\sqrt{C^2 + 4n_i^2} - C}{2}. \tag{15.37}$$

- **Interfases:**

En la mayoría de los dispositivos que se simulan existen diferentes materiales con distintas constantes dieléctricas que obligan a calcular las concentraciones y los campos en los puntos correspondientes a las interfases.
El potencial es fijado mediante la ley de Gauβ en forma diferencial:

$$\varepsilon_{sem} \left.\frac{\partial \psi}{\partial \boldsymbol{n}}\right|_{sem} - \varepsilon_{ins} \left.\frac{\partial \psi}{\partial \boldsymbol{n}}\right|_{ins} = Q_{\text{int}} \qquad (15.38)$$

donde el subíndice *sem* corresponde a las magnitudes en el semiconductor, el subíndice *ins* al aislante y Q_{int} es la carga superficial encerrada en la interfase.
En las ecuaciones de continuidad se tiene que la componente normal a la interfase debe ser igual a la tasa de recombinación superficial R^{SURF}

$$\boldsymbol{J}_{\boldsymbol{n}} \cdot \boldsymbol{n} = -qR^{SURF} \qquad (15.39)$$

$$\boldsymbol{J}_{\boldsymbol{p}} \cdot \boldsymbol{n} = qR^{SURF} \qquad (15.40)$$

- **Fronteras artificiales:**

En este caso las fronteras son impuestas bien porque el dispositivo es demasiado grande y en esa zona las variaciones de las magnitudes son muy pequeñas (por ejemplo en las zonas lejanas a la región de deplexión de un MOSFET) o por consideraciones de autocontención para el espacio a simular (zonas de drenador y fuente en las que no se aplica directamente el potencial externo y no están cubiertas por óxido).
Es posible aplicar condiciones de tipo Dirichlet o también de tipo Neumann

$$\frac{\partial \psi}{\partial \boldsymbol{n}} = 0 \qquad (15.41)$$

$$\frac{\partial n}{\partial \boldsymbol{n}} = 0 \qquad (15.42)$$

$$\frac{\partial p}{\partial \boldsymbol{n}} = 0 \qquad (15.43)$$

En cualquier caso, el uso de estas condiciones debe estar justificado previamente por razonamientos matemáticos o físicos de forma que el error introducido por la frontera artificial sea despreciable.

15.6 Simuladores Comerciales

Existen infinidad de simuladores para propósitos de investigación. La mayoría de ellos son bastante limitados en el sentido de que sólo permiten estudiar determinado tipo de estructuras o utilizan modelos bastante complejos que los aquí presentados. Sin embargo, en la industria la simulación se lleva a cabo con herramientas específicas que permiten realizar una gran cantidad de análisis de distintos tipos y tiene un entorno bastante amigable con el usuario. Estos simuladores buscan fundamentalmente optimizar los tiempos de desarrollo, por tanto, salvo en ciertas excepciones, el modelo estándar para el transporte usado en la industria es el de difusión y deriva.

A continuación se presentan algunos de los simuladores comerciales más utilizados. Algunos de ellos forman parte de un paquete completo que no solo simula el comportamiento del dispositivo, sino que parte de la simulación de los procesos, continua con el dispositivo y puede subir un nivel más para integrar estos dispositivos en circuitos.

ATLAS Device Simulation Software de la compañía Silvaco (www.silvaco.com) ofrece un paquete de simulación integrado por diferentes módulos que permite simular el comportamiento eléctrico, óptico y térmico de dispositivos semiconductores. El programa permite realizar análisis DC, AC y respuesta en el dominio del tiempo de dispositivos basados en Si y en materiales compuestos III-V y II-VI tanto en 2 como en 3 dimensiones. Los diferentes módulos disponibles dependiendo del tipo de simulación que quiera realizarse son los siguientes:

- **S-Pisces:** Simulador 2D/3D que resuelve las ecuaciones de difusión, deriva y de balance de energía para estructuras basadas en silicio. Incluye varios modelos de movilidad y generación recombinación.
- **Blaze 2D/3D:** Simulador de propósito general para materiales avanzados (III-V, II-VI, binarios, ternarios, cuaternarios). Este módulo incluye los modelos para difusión y deriva y balance de transporte de energía además de los distintos modelos usados en S-Pisces para generación-recombinación.
- **Quantum 2D/3D QCEM:** Módulo para simulación de diferentes efectos cuánticos debidos al confinamiento de los portadores en las estructuras a estudio (como en el caso de dispositivos SOI de lámina delgada o en hetroestructuras). Incluye cálculo de las energías para los distintos estados ligados y autofunciones.
- **Giga:** Giga calcula efectos térmicos locales tales como generación de calor, flujo de calor, self-heating o calentamiento de la red de una forma autoconsistente.

- **Mocasim:** Módulo que implementa un generador de parámetros utilizando el método de Ensemble Monte Carlo. Incluye mecanismos de dispersión tanto intravalle como intervalle.

DESSIS (ISE) www.ise.ch simula las características eléctricas, ópticas y térmicas de dispositivos semiconductores pudiendo describir sistemas con geometrías 1D, 2D y 3D. Sus módulos incluyen la resolución de la ecuación de transporte en su aproximación hidrodinámica, transporte cuántico, efecto túnel, transporte no local y efectos de electrones calientes.

DAMOCLES cuyo nombre viene de **D**evice **A**nalysis Using **Mo**nte **C**ar**l**o **e**t Poisson solver es un simulador desarrollado por IBM www.research.ibm.com/DAMOCLES/home.html que resuelve mediante el método Monte Carlo la BTE y Poisson de forma autoconsistente. También es posible acoplar la ecuación de Schrödinger para que los efectos cuánticos sean tenidos en cuenta. DAMOCLES utiliza una estructura de bandas completa resolviendo la BTE en 3D para el espacio de momentos y en 2D en el espacio real. El tiempo es también una variable incluida en el simulador con lo que resulta ser un simulador 6D.

Ejemplo 15.1 La ecuación de Difusión 1D

Como ejemplo de aplicación se va a desarrollar la ecuación de difusión en 1D dependiente del tiempo por el método de las diferencias finitas.

$$\frac{\partial f}{\partial t} = a\frac{\partial^2 f}{\partial x^2} \qquad (15.44)$$

Se va a estudiar la difusión de un exceso de portadores en un semiconductor generado por un pulso de luz en el centro del mismo aplicado durante un tiempo muy pequeño. La recombinación de los portadores está implícitamente despreciada en la ecuación donde *a* representa la constante de difusión de los portadores en el medio.

Para utilizar el formalismo de diferencias finitas es necesario realizar una discretización tanto espacial como temporal. Para ello se dividirá el espacio y el tiempo en intervalos de igual amplitud, Δx y Δt respectivamente.

Para la aproximación de la primera derivada temporal se utilizará el siguiente esquema:

$$\frac{\partial f}{\partial t} = \frac{f^{n+1}(x) - f^{n}(x)}{\Delta t} \qquad (15.45)$$

donde los superíndices indican la variable temporal discreta de forma que $f^{n}(x) = f(x, n\Delta t)$. Aplicando la expresión (15.30) para la derivada segunda espacial se tiene finalmente:

$$\frac{f^{n+1}(x)-f^{n}(x)}{\Delta t}=a\frac{f^{n}(x+\Delta x)-2f^{n}(x)+f^{n}(x-\Delta x)}{\Delta x^{2}} \qquad (15.46)$$

Con lo que finalmente se puede obtener una relación recurrente para calcular la evolución espacial y temporal de la distribución de portadores:

$$f^{n+1}(x)=f^{n}(x)+a\Delta t\frac{f^{n}(x+\Delta x)-2f^{n}(x)+f^{n}(x-\Delta x)}{\Delta x^{2}}. \qquad (15.47)$$

Se puede demostrar que para que la solución sea estable y converja es necesario que se cumpla la siguiente relación entre los intervalos espaciales y temporales:

$$\Delta t\le\frac{(\Delta x)^{2}}{2a}. \qquad (15.48)$$

En la Figura 15.6.1 se observa la evolución espacial y temporal cuando se generan en un intervalo *dt* 10^7 portadores. Se impone como condiciones de contorno que la concentración de portadores se anule en los extremos del semiconductor y se toman $\Delta t=10^{-7}$, $\Delta x=10^{-3}$ y a = 1, todo medido en unidades arbitrarias.

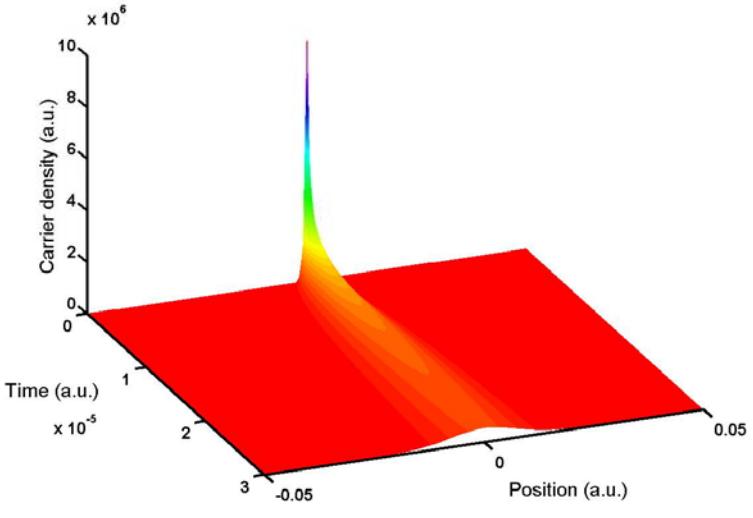


Figura 15.6.1 Solución a la ecuación de difusión monodimensional dependiente del tiempo

Por último en la Figura 15.6.2 se puede observar como, a diferencia del caso anterior, si el espaciado del mallado no se ajusta correctamente el resultado oscila y puede terminar divergiendo dando un resultado que no es aceptable físicamente.

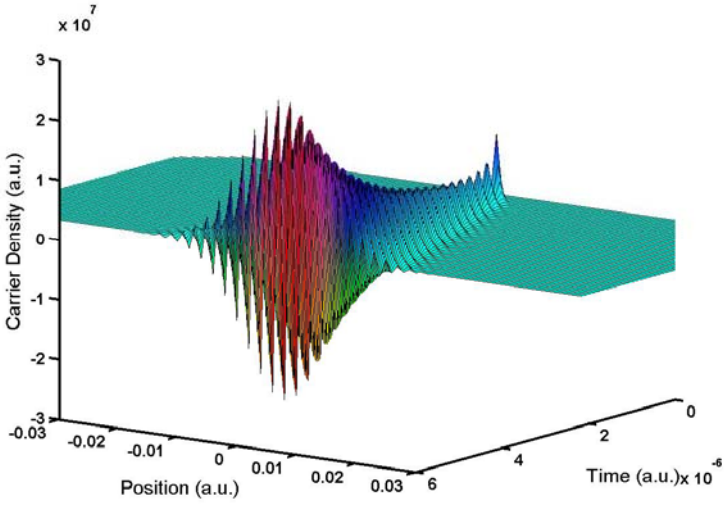


Figura 15.6.2 Resultado obtenido para un $\Delta t = 5.2\times10^{-7}$ que no cumple la condición (15.48) donde se observan oscilaciones en la solución.

Simulación

A continuación, a modo de ejemplo, se presenta el código que genera la solución del ejemplo anterior utilizando MATLAB

```
% Resolucion de la ecuacion de difusion por el metodo de
% finitas diferencias

% Definicion de los parametros en unidades arbitrarias

dt= 1e-7;   % Incremento de tiempo
dx= 1e-3;   % Incremento espacial
dif= 1;     % Coeficiente de difusion


% Definicion del grid espacio temporal

sol=zeros(101,300);     % Matriz de la solucion cada columna
                         % corresponde a la distribucion de
                         % portadores en un momento dado
sol(51,1)= 1e7;       % Condiciones iniciales de portadores

% Resolucion de la ecuacion

for i=2:300
    for j=2:100
        sol(j,i)=(dif*dt/(dx*dx))*(sol(j+1,i-1)
        -2*sol(j,i-1)+sol(j-1,i-1))+sol(j,i-1);
    end
end
```

RESUMEN

En este capítulo se han presentado los conceptos básicos relacionados con la simulación de dispositivos. Se ha justificado la necesidad de realizar simulaciones tanto en el campo industrial como en el de investigación. Se han presentado diferentes tipos de simuladores en función de la aproximación utilizada para describir el transporte de los portadores describiéndose los modelos que utilizan cada uno de ellos. Se han estudiado diferentes modelos de movilidad y de generación-recombinación de portadores que permiten ampliar el rango de validez de las ecuaciones. Se ha tratado el problema de la discretización tanto del espacio de simulación como de las ecuaciones que describen el sistema. Para terminar, se han enumerado distintas herramientas de simulación que pueden ser encontradas en el mercado.

REFERENCIAS

[1] http://silvaco.com/products/TCAD.html

[2] **www.research.ibm.com/DAMOCLES/home.html**

[3] **www.ise.ch**

[4] M. Heath. *Scientific Computing: an Introductory Survey, 2nd Ed.* McGraw-Hill, NewYork, 2002

[5] U. Ravaioli. *ECE439 - Advanced Theory of Semiconductors and Semiconductor Devices Numerical Methods and Simulation.* 1994

[6] S. Selberherr. *Analysis and Simulation of Semiconductor Devices.* Springer-Verlag, Wien, 1984

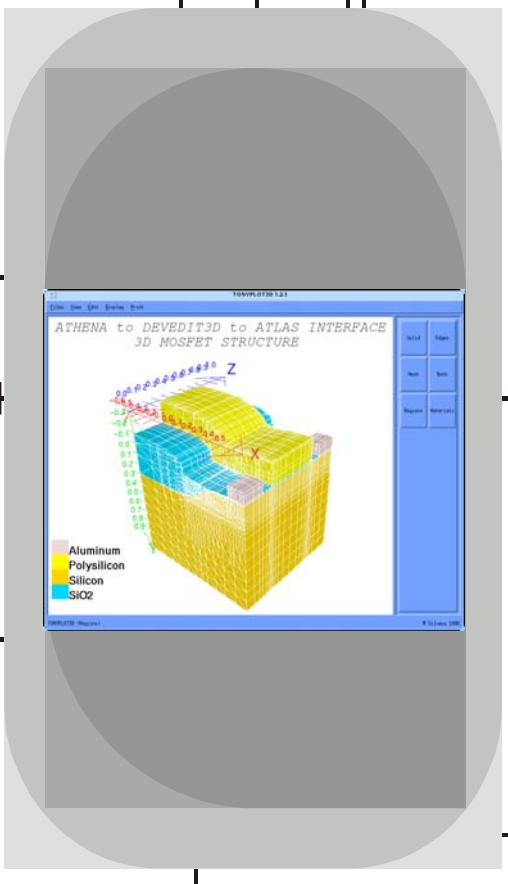
[7] S. Wolf. *Silicon Processing for the VLSI Era, Vol 3: The Submicron MOSFET*, Lattice Press, Sunset Beach, 1995

16

Capítulo

Modelos para simulación de procesos

Simulación



Índice

Objetivos

- Presentación de los modelos utilizados por los simuladores de procesos.
- Describir los modelos matemáticos de difusión de impurezas y defectos.
- Descripción matemática del proceso de oxidación del silicio y del polisilicio.
- Formación de siliciuros en los contactos.
- Descripción de los modelos analíticos y estadísticos de implantación iónica en materiales cristalinos y amorfos.
- Comentar brevemente la capacidad de los simuladores para reproducir el grabado y la difusión en SiGe.
- Simplificar los modelos equivalentes encontrados para facilitar el análisis de estos circuitos.

Palabras Clave

Difusión
Silicidación
Oxidación
Implantación Iónica
Grabado (Etching)
Deposición
SiGe
Epitaxia
Materiales amorfos y cristalinos
Athena, SSuprem4
Simulación Monte Carlo
Ley de Fick
RTA
ETD, OED
Distribuciones Gaussiana, Pearson y doble Pearson
Modelo CNET de difusión
Stress
Clusters
Oxidación lineal - parabólica
Vacantes, intersticiales
Modelos de Fermi y bidimensional

16.1 Importancia de la simulación de procesos

Simulación de Procesos
Permite fijar una secuencia de procesos del tipo oxidación, implantación iónica, difusión, grabado, crecimiento epitaxial, así como establecer la duración temporal y la temperatura a la que se produce cada uno de ellos para poder predecir la estructura resultante.

Los simuladores de procesos basados en modelos físicos predicen la estructura resultante de la secuencia de procesos tecnológicos especificados. Esto se realiza resolviendo el sistema de ecuaciones que describen los procesos físico-químicos. Estos simuladores proporcionan tres ventajas fundamentales: Predicen la estructura resultante, permiten comprender qué sucede y hacen transparente todo el conocimiento teórico necesario para su realización a usuarios inexpertos en el tema.

La simulación basada en principios físicos es diferente al modelado empírico. El objetivo del modelado empírico es obtener una fórmula analítica que se ajuste a los datos existentes con precisión y complejidad mínima proporcionando una aproximación útil y eficiente. El modelado empírico proporciona representaciones compactas de datos proporcionados por cualquier fuente. Como contrapartida, no aporta una mejor comprensión del proceso y carece de capacidades predictivas.

Las simulaciones basadas en fundamentos físicos son una alternativa a los experimentos como fuente de información y son importantes por dos razones: 1.- Casi siempre son mucho más rápidas y baratas que diseñar e implementar que los experimentos. 2.- Proporcionan información que es difícil o imposible de medir. El principal inconveniente es que toda la física/química relevante del proceso debe incorporarse al simulador y para ello es necesario desarrollar los métodos numéricos necesarios para resolver el sistema de ecuaciones resultante. El usuario del simulador debe especificar:

• La geometría de la estructura a simular.

• La secuencia de procesos (implantación, grabado, difusión, etc) que se van a simular.

• Los modelos físicos a utilizar.

A continuación se presentan, en cada uno de los apartados de este capítulo, los modelos que los simuladores comerciales utilizan para reproducir el resultado de la oxidación, silicidación, implantación iónica y algunos otros procesos.

16.2 Modelos de Difusión

Los modelos de difusión describen la manera en que los perfiles implantados de dopantes y/o defectos se redistribuyen durante los ciclos térmicos debido a los gradientes de concentración y a los campos eléctricos internos. Al modelar el proceso de difusión, hay efectos adicionales a considerar, por ejemplo, la formación de clusters de impurezas, activación de las mismas y el tratamiento de las interfases.

La difusión de dopantes y defectos puntuales se describe mediante diferentes modelos. Los tres más utilizados son:

- El modelo de difusión de Fermi.
- El modelo de difusión en dos dimensiones.

- El modelo acoplado de difusión.

Cada modelo es una extensión del anterior. Así el de Fermi está incluido en el de dos dimensiones, que a su vez se incluye en el modelo acoplado. Las dos diferencias significativas entre ellos son la manera en que se tratan los defectos puntuales en la simulación, y cómo se formulan las difusividades específicas de cada impureza dopante.

Los tres modelos se basan en el concepto de la difusión del par, que dice que un átomo dopante no puede difundirse por sí solo, necesita la presencia de un defecto puntual (intersticial o vacante) en su entorno próximo como vehículo de la difusión. Si existe una energía de enlace no despreciable entre los dos, pueden moverse como una entidad (un par). Cuando se habla de la difusividad del dopante se está haciendo mención a la difusividad del par en su conjunto. Un defecto puntual, sin embargo, puede difundirse libremente o como participante en un par dopante-defecto. La difusividad de un defecto puntual libre, puede ser diferente a la difusividad de un par dopante-defecto.

Todos los modelos de difusión empleados en simuladores utilizan los conceptos de concentración química y concentración activa. La concentración química hace referencia a la cantidad total de dopante implantado. Sin embargo, cuando los dopantes están presentes en elevadas concentraciones se produce la formación de clusters o la desactivación eléctrica de manera que la concentración eléctricamente activa es menor que la concentración química correspondiente.

Descripción matemática

La definición matemática de un modelo de difusión incluye las siguientes especificaciones:

- Una ecuación de continuidad (ecuación de difusión).
- Uno o más términos de flujo.
- Un conjunto de condiciones de contorno.

En el caso de la difusión de impurezas en semiconductores, se necesita una ecuación para cada impureza dopante presente y para cada tipo de defecto si éstos se representan explícitamente en el modelo. Puesto que los dopantes se difunden únicamente como miembros de pares dopante-defecto, la ecuación de continuidad del dopante es realmente una ecuación de continuidad para las parejas defecto-dopante.

La formulación de las ecuaciones de continuidad se realiza partiendo de un número de suposiciones:

- Los procesos electrónicos ocurren en una escala temporal mucho menor que la del resto de los procesos (aproximación adiabática).
- La reacción de emparejamiento entre dopantes y defectos se supone siempre en equilibrio.
- Los dopantes móviles son eléctricamente activos y viceversa.

Ecuación de difusión genérica

Todos los modelos de difusión, siguen la misma formulación matemática genérica de la ecuación de continuidad. Ésta expresa la conservación de partículas, es decir, la velocidad de cambio con el tiempo del número de partículas en un volumen unidad debe ser igual al número de partículas que abandonan ese volumen debido a la difusión, más el número de las partículas que se crean o aniquilan en ese volumen debido a los posibles términos fuente y sumidero. Esta ecuación básica de continuidad para la difusión de un tipo concreto de partículas (C) en un cierto volumen del semiconductor es simplemente una ecuación de Fick de segundo orden

$$\frac{\partial C_C}{\partial t} = -\nabla J_A + S \ , \qquad (16.1)$$

donde C_C es la concentración total de partículas, J_A el flujo de partículas móviles y S representa los términos de fuente y sumidero.

En problemas de difusión en un semiconductor hay generalmente dos contribuciones al flujo de partículas. La primera es proporcional al gradiente de la concentración de partículas móviles y el factor de proporcionalidad, D_A, se conoce como coeficiente de difusión o difusividad. La segunda es un término de deriva, proporcional al campo eléctrico local

$$J_A = -D_A(C)\nabla C_A + C_A \mu E \ , \qquad (16.2)$$

donde C_A es la concentración de impurezas móviles, μ es la movilidad y E el campo eléctrico. En equilibrio termodinámico se cumple la relación de Einstein que relaciona la movilidad con el coeficiente de Difusión a través de la relación $D = \frac{kT}{q}\mu$.

Modelo de Fermi

El modelo de Fermi supone que las concentraciones de defectos puntuales están en equilibrio termodinámico y no necesitan una representación directa. El efecto de la presencia de defectos puntuales sobre la difusión de dopantes se tiene en cuenta en las difusividades del par impureza-defecto. La ventaja principal de usar el modelo de difusión de Fermi es que la velocidad de la simulación aumenta considerablemente ya que los defectos puntuales no se representan directamente y sólo se necesita simular la difusión de dopantes, de manera que el problema numérico se simplifica considerablemente. Pero, puesto que los defectos no se simulan directamente, el modelo de Fermi no puede reproducir de manera fiable ciertas situaciones en las cuales las poblaciones de defectos no están en equilibrio (por ejemplo en la oxidación húmeda o en las difusiones de emisor-base).

Modelo bidimensional

En este modelo, las poblaciones de defectos puntuales se representan explícitamente y se analiza su evolución temporal. Si se produce una saturación de defectos, la difusividad del dopante se ve afectada mediante un factor de escala. Por lo tanto, con perfiles de defectos en equilibrio, el modelo bidimensional se reduce al modelo de Fermi, aunque no es tan rápido en este caso ya que debe resolver las ecuaciones de los defectos que realmente no son necesarias. El acoplamiento entre defectos y dopantes en este modelo es unidireccional. Esto significa que la difusión de dopantes está muy influenciada por la difusión de los defectos puntuales, mientras que la difusión de defectos se considera totalmente independiente de la difusión de dopantes. En términos físicos, esto corresponde a un apareamiento entre defectos y dopantes con una energía de enlace nula.

La diferencia principal entre el modelo de Fermi y el modelo de dos dimensiones es la representación y la evolución explícitas de las poblaciones de defectos fuera del equilibrio. Por lo tanto, hay tres ecuaciones de difusión que gobiernan el proceso: una para los dopantes, otra para los defectos intersticiales y una última para los defectos vacantes. Además, también es necesario tener cuenta la formación y disolución de clusters < 311 >, la generación de defectos puntuales producidos por la oxidación y la recombinación en el volumen y en las interfases.

Algunos comentarios sobre la difusión de defectos

Los defectos puntuales tienen difusividades mayores que los dopantes y pueden difundirse a mayores profundidades en la estructura durante la simulación. Si la geometría utilizada es poco profunda, podemos obtener resultados carentes de significado físico como por ejemplo una elevada concentración de defectos en regiones donde los dopantes están presentes y como consecuencia un aumento ficticio de su difusividad. Por lo tanto, es posible que necesitemos ampliar la profundidad de la estructura simulada para proporcionar un sumidero adecuado a los defectos puntuales. Para determinar esta profundidad debemos estimar la longitud de difusión característica del defecto usando la siguiente expresión

$$l = \sqrt{D_x \Delta t} \tag{16.3}$$

donde D_x es el coeficiente de difusión del defecto y Δt el tiempo total de difusión. Las simulaciones demuestran que una profundidad de entre 20 y 50 micras es suficiente en la mayoría de los casos. Esta restricción en la profundidad mínima de la estructura plantea una amenaza a la eficacia del cálculo siempre que se empleen los modelos de difusión que incluyen defectos puntuales. No obstante, puesto que no se necesita una gran precisión en el perfil de defectos cerca del fondo de la estructura,

podemos reducir el coste de cómputo haciendo el grid más grueso en esa región.

Modelo de acoplamiento total

El modelo completamente acoplado de difusión es idéntico al modelo de dos dimensiones. La única diferencia importante es que la difusión de los defectos ahora está influenciada por la difusión de los dopantes mediante la inclusión de los flujos de pares impureza-defecto al término de flujo en la ecuación de los defectos. Así, ahora existe una verdadera interacción en dos direcciones entre la difusión de dopantes y la difusión de defectos. El modelo acoplado es ligeramente más costoso computacionalmente que el modelo de dos dimensiones, pero incluye la capacidad de reproducir ciertos aspectos importantes del procesado de semiconductores. Sin embargo, desde un punto de vista físico este modelo es incapaz de representar explícitamente pares dopantes-defectos y una subdivisión clara de defectos y de dopantes en fracciones apareadas y no-apareadas. Por lo tanto, este modelo no puede reproducir la saturación de la difusividad del dopante que se piensa que ocurre cuando todos los dopantes están apareados con defectos (por ejemplo en zonas muy dañadas por implantación). En otras palabras, el modelo confía en la aproximación diluida, la cual supone que la concentración de pares es mucho más pequeña que la de dopantes y defectos.

Modelado de difusión RTA

SSUPREM4 posee la capacidad de modelar procesos térmicos rápidos de recocido (Rapid Termal Annealing, RTA) en el marco de los modelos existentes de difusión (es decir, el modelo de dos dimensiones y el modelo completamente acoplado). Puesto que RTA es básicamente un ciclo térmico breve consistente en una elevación rápida de las temperaturas, el fenómeno transitorio del realce de la difusión (TED) dominará siempre que se produzca una cantidad apreciable de daños en la estructura cristalina. Debido a que la difusión del dopante está muy relacionada con la evolución de las poblaciones de defectos podemos calibrar estos modelos a las condiciones de RTA mediante un ajuste de los parámetros relacionados con defectos puntuales.

Modelos de desactivación eléctrica y formación de clusters

Cuando los dopantes están presentes en elevadas concentraciones, la concentración (móvil) eléctricamente activa, C_A , puede ser menor que la concentración química correspondiente, C_C . Para que una impureza llegue a ser eléctricamente activa en un material semiconductor, debe incorporarse a la red sustituyendo a un átomo, en cuyo caso contribuirá con un portador a la Banda de Valencia (impurezas aceptadoras) o a la Banda de Conducción (impureza donadora). Pero por encima de ciertas concentraciones, no es posible incorporar más dopantes

como sustitutos de átomos semiconductores. El exceso resultante será inactivo. El umbral en que se produce esta desactivación se suele llamar el límite sólido de la solubilidad, que es una terminología algo imprecisa puesto que las impurezas pueden existir en diversas fases en el cristal. Por lo tanto, no está bien definido a qué transición de fase se refiere el límite sólido de la solubilidad. Por ejemplo, los dopantes en exceso podrían participar en pequeños clusters o en precipitados más grandes. El umbral de desactivación sería una designación más apropiada para este límite.

Modelo de activación eléctrica

El objetivo del modelo de activación eléctrica es calcular la concentración de dopante en la que se produce la desactivación. Para ello se utilizan dos propuestas diferentes:

1.- El modelo de clustering As-vacante para el arsénico se basa en la reacción química:

$$As^{+} + V^{=} \rightleftarrows AsV^{-} \tag{16.4}$$

2.- El modelo de activación semiempírico tabulado se utiliza para todos los dopantes excepto el arsénico. El programa interpola en una tabla de datos experimentales que contiene la temperatura y el correspondiente umbral de desactivación calculando el valor C_A^{th} que corresponde a la temperatura de simulación.

Modelo de activación transitoria

El modelo transitorio de activación supone que los dopantes justamente después de ser implantados son inactivos. Después de que un ión se implanta en el silicio, se requiere cierto tiempo antes de que lleguen a ser eléctricamente activos. Por tanto, este modelo asume que todos los dopantes son inicialmente inactivos y puede que no se activen inmediatamente sino que llegan a serlo gradualmente después de cierto tiempo. El modelo transitorio de activación simula este comportamiento y se aplica para activar los dopantes implantados. Para ello debemos resolver la siguiente ecuación para la concentración activa C_A:

$$\frac{\partial\left(C_C - C_A\right)}{\partial t} = \frac{C_C - C_A^{eq}}{\tau_A} \tag{16.5}$$

donde C_C es la concentración química del dopante y C_A^{eq} es la concentración activa en equilibrio. El parámetro τ_A , conocido como constante de tiempo para la activación, es una función de la temperatura y se calcula mediante una expresión de tipo Arrhenius.

$$\tau_A = \tau_{A0} \cdot \exp\left(-\frac{TRACT.E}{kT}\right) \tag{16.6}$$

Modelo CNET de difusión para elevadas concentraciones

El objetivo principal de la simulación es calcular las características eléctricas de un dispositivo usando exclusivamente como parámetros de entrada información de los procesos tecnológicos utilizados en su fabricación. Puesto que las características eléctricas del dispositivo dependen fundamentalmente de la distribución de impurezas, es importante que los modelos de difusión utilizados en la simulación de procesos sean lo más exactos posible. Esto es particularmente importante para los dispositivos de dimensiones submicra.

Las distribuciones bidimensionales de dopantes activos sólo pueden obtenerse mediante simulación basados en diferentes modelos. Está comprobado que el comportamiento anormal de la difusión de dopantes en silicio está producido por defectos puntuales fuera del equilibrio. El modelo de difusión CNET fue desarrollado en CNET-Grenoble (France Telecom)

Los principales aspectos físicos incluidos en el modelo son:

- La difusión de dopantes es asistida por vacantes (V) e intersticiales (I). Estos defectos puntuales se pueden encontrar en distintos estados de carga y las concentraciones relativas de cada uno dependen de la posición del nivel de Fermi (es decir de la concentración local de dopante).
- Los dos, V e I, presentan una elevada energía de enlace con los átomos del dopante, y por lo tanto, las especies que se difunden son pares dopante-defecto. Se puede concluir que los dopantes aislados son inmóviles. Los pares impureza/defecto, en sus diferentes estados de carga, se supone que se encuentran en equilibrio local con los átomos sustitutivos libres del dopante y los defectos libres.
- En el caso de As y B en concentraciones próximas al límite de la solubilidad en sólidos, se forman complejos neutros e inmóviles (As_2V o B_2I), que disminuyen la difusividad efectiva y contribuyen a aumentar la concentración de dopantes eléctricamente inactivos.
- Cuando la concentración de dopante excede de 10^{20} cm^{-3}, los pares dopante-vacante no pueden considerarse como entidades aisladas porque las vacantes pueden interaccionar con más de un átomo del dopante. En el modelo CNET, este hecho se describe mediante un cluster de átomos del dopante donde la difusividad eficaz y la concentración de las vacantes aumentan fuertemente, causando un realce de la difusión asistida por vacantes.
- El flujo de cada especie que se difunde (pares dopante/defecto y defectos libres) incluye los términos de deriva causados por el campo eléctrico producido por los gradientes del dopante.

- Intersticilaes y Vacantes no se consideran en equilibrio local, pero pueden ser aniquilados mediante un proceso de recombinación bimolecular.

Modelo de difusión en polisilicio

El mecanismo para la difusión de impurezas en polisilicio es diferente que en el silicio cristalino. En comparación con las regiones de interés en el dispositivo, el polisilicio se caracteriza por una microestructura de regiones cristalinas pequeñas denominadas granos. Estos granos se encuentran separados en sus contornos que ocupan cierto volumen espacial y se conectan para formar una red compleja. La textura y la morfología de la estructura del grano dependen de las condiciones de la deposición y del tratamiento térmico posterior. Las impurezas dentro del grano se difunden de manera diferente que en la frontera con otros cristales de polisilicio. La simulación completa de la difusión dentro de todas las regiones del polisilicio con la complejidad geométrica que ello conlleva, es demasiado costosa en tiempo de cómputo y por lo tanto requiere un tratamiento matemático especial.

SSUPREM4 incorpora un modelo numérico de dos dimensiones para la difusión de la impureza en polisilicio. En este modelo, se describe matemáticamente la microestructura del polisilicio usando una aproximación de homogeneización local.

En esta aproximación, cada cristal de polisilicio está formado por dos componentes, el interior y el contorno del cristal. La red formada por los contornos de los granos se caracteriza por una función escalar que describe el tamaño del grano y una función vectorial que describe la dirección del contorno. En consecuencia, la difusión de cada impureza se divide en dos componentes: dentro del grano y en el contorno del grano. Durante un tratamiento térmico, la recristalización del polisilicio también se modela para incluir el crecimiento del tamaño de los cristales.

16.3 Modelos de Oxidación

La fabricación de circuitos integrados depende esencialmente del proceso de oxidación térmica para la formación de dieléctricos de puerta, de las regiones de aislamiento del dispositivo, de las regiones spacer y de las regiones de máscara de la implantación iónica. El control exacto del espesor del dióxido del silicio resulta ser de máxima importancia a medida que los geometrías del dispositivo continúan reduciéndose a dimensiones nanométricas. La oxidación se produce cuando el silicio (o el polisilicio) se expone a un ambiente oxidante. La simulación de la oxidación del polisilicio se realiza de una manera muy similar a la del silicio.

Normalmente las superficies expuestas del silicio tienen una fina capa de óxido. SSUPREM4 deposita automáticamente una fina capa de óxido en todas las superficies expuestas del silicio (polisilicio) al principio del proceso de oxidación con un valor por defecto de 20 Å. Los

modelos bidimensionales de oxidación se basan en la teoría lineal-parabólica de Deal y Grove [Deal 1965] en la que la oxidación del silicio se modela considerando tres procesos:

(1) El oxidante (H_2O, O_2) se transporta desde el gas ambiente al interior del SiO_2 a través de la interfase gas/SiO_2.

(2) El oxidante se transporta a través de la capa de SiO_2 hasta alcanzar la interfase Si/SiO_2.

(3) El oxidante, al alcanzar la interfase Si/SiO_2, reacciona con silicio para formar una nueva capa de SiO_2.

El transporte del oxidante a través de la interfase gas/SiO_2 se representa mediante:

$$\boldsymbol{F_1} = h\left(C^* - C_O\right)\boldsymbol{n_o} \tag{16.7}$$

donde h es el coeficiente de transporte de masa en fase gaseosa, C^* es la concentración en equilibrio del oxidante en el SiO_2, C_0 es la concentración del oxidante en el óxido SiO_2 en la interfase gas/SiO_2 y $\boldsymbol{n_0}$ es un vector unitario perpendicular a la interfase gas/SiO_2 que señala hacia la capa del silicio. La concentración en equilibrio del oxidante en SiO_2 mantiene una relación lineal con la presión parcial del oxidante, P, en el gas por la ley de Henry

$$C^* = K \cdot P \tag{16.8}$$

donde K es una constante. La difusión de las moléculas del oxidante en el SiO_2 es producida por un gradiente de la concentración y expresada mediante la ley de Ficks como:

$$F_2 = -D_{eff}\nabla C \tag{16.9}$$

donde D_{eff} es el coeficiente de difusión del oxidante (O_2 o H_2O) en la lámina de óxido.

Modelos numéricos de oxidación

En la sección previa, se ha presentado una introducción al modelado unidimensional de la oxidación. A continuación se tratará de describir los modelos bidimensionales.

Los modelos numéricos de oxidación necesitan resolver la ecuación de difusión del oxidante en incrementos temporales en los puntos del mallado contenidos en la capa de SiO_2 conforme este aumenta de tamaño. La ecuación de la difusión del oxidante viene dada por:

$$\frac{\partial C}{\partial t} = \nabla \cdot F \tag{16.10}$$

donde C es la concentración de oxidante en el SiO_2, t es el tiempo de la oxidación y F es el flujo del oxidante. La ecuación (16.10) se resuelve sustituyendo la expresión (16.9) para F, y definiendo condiciones de contorno apropiadas en las interfases con el SiO_2.

La ecuación (16.10) es suficiente para describir el movimiento de la interfase Si-SiO_2 siempre que el flujo del óxido se realice en la misma dirección que el crecimiento (por ejemplo: oxidación en superficies planas). En la mayoría de las estructuras de interés, el flujo del óxido es en dos dimensiones. Por tanto, se deben incluir otras ecuaciones que tengan en cuenta este hecho. El método más común consiste en resolver una ecuación hidrodinámica simplificada del flujo utilizando los modelos de compresión o viscoso para calcular el flujo bidimensional de los elementos óxido.

La ecuación hidrodinámica a resolver es:

$$\mu \nabla^2 V = \nabla P \tag{16.11}$$

donde P es la presión hidrostática, V es la velocidad del oxidante y μ la viscosidad del óxido.

Relación lineal

Para tiempos breves y bajas temperaturas de oxidación, el crecimiento del óxido mantiene una relación lineal con el tiempo de oxidación. Los procesos que tienen lugar en las interfases (transporte del oxidante a través de la interfase gas/SiO_2 y del oxidante en la interfase Si/SiO_2) constituyen un factor determinante en la descripción de la cinética del crecimiento. En este régimen el espesor del óxido se puede aproximar como:

$$x_0 \cong (B/A) \cdot t \tag{16.12}$$

donde (B/A) es la constante de la relación lineal.

Este factor de crecimiento lineal depende de diferentes parámetros físicos que será comentados a continuación.

Se sabe que la orientación del substrato afecta la cinética de oxidación. La influencia de la orientación sobre la constante de crecimiento se modela mediante $(B/A)_{ori}$ (p. ej. $(B/A)_{ori}$ =1 para substratos <111>)

Utilizando valores elevados de la presión durante la oxidación se pueden hacer crecer láminas gruesas de SiO_2 manteniendo una temperatura baja de manera que se evite la redistribución de dopantes. La dependencia con la presión viene expresada como

$$\left(\frac{B}{A}\right)_P = P^{L.PDEP} \tag{16.13}$$

donde P es la presión parcial del gas oxidante.

El uso de cloro durante la oxidación consigue mejorar la pasivación y aumentar la resistencia dieléctrica del óxido. Para oxidaciones en ambiente seco, el HCl reacciona con el O_2 y produce H_2O y Cl_2 como productos. Para modelar la constante de crecimiento $(B/A)_{Cl}$ se utiliza un modelo tabulado como función de la temperatura y del porcentaje de HCl.

La formación de SiO_2 en substratos con elevada concentración de impurezas dopantes tipo p o n se ve favorecida en comparación con la producida en substratos cuasi-intrínsecos. Esta dependencia se manifiesta también en la constante de crecimiento $(B/A)_{\text{doping}}$ y el efecto de los elevados dopados debido al efecto eléctrico.

Finalmente, teniendo en cuenta todos estos factores, la constante se puede escribir como

$$\left(\frac{B}{A}\right)_{total} = \left(\frac{B}{A}\right)_{i}\left(\frac{B}{A}\right)_{ori}\left(\frac{B}{A}\right)_{P}\left(\frac{B}{A}\right)_{Cl}\left(\frac{B}{A}\right)_{doping} \tag{16.14}$$

Relación parabólica

Para periodos largos de oxidación y altas temperaturas el crecimiento del óxido mantiene una dependencia parabólica con el tiempo de oxidación. La difusión del oxidante en el óxido es el factor determinante en la descripción de la cinética del crecimiento. En estas condiciones el grosor del óxido se puede aproximar como:

$$x_0^2 \cong Bt \tag{16.15}$$

donde B es la constante en la relación parabólica.

Se ha comprobado experimentalmente que la constante B depende de factores tales como la presión ambiental o el contenido de HCl durante la oxidación, y se expresa como:

$$B = B_i \cdot B_P \cdot B_{HCl} \tag{16.16}$$

El término B_i se determina en función de la temperatura y del tiempo de oxidación para los substratos ligeramente dopados recocidos a presión atmosférica y sin la presencia de HCl en el ambiente.

La constante B varía con la presión ya que depende de la concentración en el equilibrio del oxidante en el óxido, C^*, que es directamente proporcional a la presión parcial del gas que oxida. La relación siguiente se utiliza para modelar esta dependencia

$$B_P = P^{P.DEP} \tag{16.17}$$

donde P es la presión parcial del gas que oxida expresada en atmósferas.

Se ha observado que la presencia de HCl durante la oxidación también afecta la velocidad de crecimiento del óxido. Diferentes estudios suponen que el HCl provoca una tensión en la estructura cristalina del óxido que a su vez aumenta el coeficiente de difusión del oxidante. La dependencia de la constante B con la concentración de HCl se modela de una manera similar al caso anterior de dependencia lineal con el tiempo de oxidación. El procedimiento es muy sencillo, dada una concentración de HCl, se utiliza una relación previamente tabulada para determinar el factor de realce que modifica la constante parabólica.

En la práctica, el gas oxidante está formado por una mezcla de gases compuesto por más de un oxidante y otras impurezas. La velocidad final de oxidación será el efecto combinado de todas estas especies. Para simular la oxidación en un ambiente de gases mezclados se calcula simultáneamente la difusión y la oxidación de cada uno de los gases que forman la mezcla.

Recomendaciones para conseguir simulaciones de oxidación correctas

Durante el proceso de simulación es normal encontrar dificultades y resultados que carecen de significado físico. Uno de los errores más comunes en la simulación de la oxidación es el de utilizar un mallado incorrecto en el óxido que puede dar lugar a una distribución de óxido en forma de dientes de sierra y a errores en la resolución de la impureza. Mientras que la capa del óxido está creciendo, se van añadiendo puntos del grid en los espaciamientos previamente definidos. Al consumirse el silicio, los dopantes se mueven a través de la interfase Si/SiO_2. Es esencial trabajar con un mallado apropiado en el óxido para explicar correctamente la redistribución del dopante y las impurezas durante el proceso de oxidación.

Un ejemplo típico donde es importante utilizar un grid fino y preciso es durante el crecimiento del oxido de puerta de un MOSFET con una puerta de polisilicio con elevada concentración de dopado. Por defecto, SSUPREM4 utiliza un espaciado en el grid de 0.1 μm. De esta forma, se añade un nuevo punto al grid del óxido cuendo este crezca un espesor de 0.1 μm (1000 Å). Esta distancia entre puntos del grid es apropiada para óxidos de campo.

Sin embargo, usar este espaciado para la formación del óxido de puerta en la tecnología MOS actual da lugar a la ausencia de puntos del grid en el interior del SiO_2. Sin un mallado fino en el óxido de puerta, al resolver la ecuación de difusión posteriormente, nos encontramos con que el dopante del polisilicio puede penetrar en el substrato subyacente del silicio. Este efecto de la simulación conduce a valores de la tensión umbral muy diferentes de los esperados.

Con frecuencia los dopantes se implantan sobre capas de óxido crecidas térmicamente. Existen dos razones importantes por las cuales es necesario tener un espaciamiento apropiado en el grid del óxido a través del cual el dopante se implanta. En primer lugar, ésto ayudará en la determinación correcta del perfil de dopante en la capa de óxido y el silicio subyacente. En segundo lugar, es necesario un grid apropiado para simular la difusión del dopante en el óxido durante los procesos posteriores. Durante el recocido el dopante se difunde en el SiO_2 y el silicio, e incluso puede llegar a evaporarse en el ambiente a través de la interfase gas/SiO2. Si el grid en el óxido no es el apropiado, la cantidad de dopante evaporado se puede subestimar, resultando una concentración de dopante conservada en el substrato mayor de la real. El origen del

problema es similar al descrito en las secciones anteriores, puede que no existan suficientes puntos del mallado en el interior del óxido. La solución se consigue agregando más puntos en el mallado del SiO_2 conforme este va creciendo.

Oxidation Enhanced Difusión (OED)

Durante la oxidación térmica del silicio parte de las impurezas presentes en el silicio se incorporan a la capa cada vez más gruesa de SiO_2. Mientras se produce la oxidación, los átomos de silicio ocupan posiciones intersticiales (los intersticiales se inyectan en el silicio desde la interfase Si/SiO_2) debido a que las moléculas de oxígeno se incorporan a la red cristalina para formar SiO_2. Como consecuencia de la inyección de defectos intersticiales durante la oxidación, es posible que el coeficiente de difusión del dopante aumente. Para que este fenómeno aparezca reflejado en el resultado final, debemos incluir la creación y el movimiento de intersticiales y vacantes en la simulación.

También puede producirse una disminución de la difusión durante la oxidación térmica. Para aquellos dopantes que se difunden sobre todo utilizando vacantes, sus difusividades se pueden reducir durante la oxidación debido a la recombinación de vacantes con intersticiales inyectados desde la interfase de SiO_2/Si.

16.4 Modelos de Silicidación

Los siliciuros se forman cuando un metal reacciona con silicio o polisilicio para crear una fase intermedia. La conductividad de los siliciuros es normalmente varios órdenes de magnitud mayor que la de las regiones altamente dopadas ($n+$ y $p+$). Las tecnologías ULSI utilizan siliciuros para reducir la resistencia de contactos e interconexiones

El crecimiento del siliciuro se describe como el movimiento de las interfases metal-siliciuro y silicio/poly-siliciuro donde los átomos del silicio y del metal reaccionan para formar el siliciuro. Se calculan las velocidades con las que se desplazan las interfases usando los coeficientes de reacción en la interfase y las concentraciones del silicio y del metal en el siliciuro. Existen diferentes combinaciones posibles de materiales: el disiliciuro de platino ($PtSi_2$), de titanio ($TiSi_2$), y el siliciuro de tungsteno (WSi_2). La silicidación se consigue mediante la deposición de capas del metal en la superficie expuesta del silicio/poly y un posterior ciclo térmico. Es necesario especificar diferentes parámetros para ajustar los coeficientes de la velocidad de reacción y el aumento o reducción de volumen. También es necesario especificar los parámetros de la difusión de átomos de silicio y del metal, así como los de tensión mecánica del metal y del siliciuro. El modelado del proceso de silicidación es similar al de oxidación. Durante cada intervalo temporal, se calculan para cada punto del grid las velocidades de crecimiento (dependientes de la temperatura), los coeficientes de reacción en la superficie y las concentraciones de silicio y metal.

La difusión del silicio y el metal dentro de la lámina de siliciuro se modela como un proceso de difusión de defectos puntuales, donde el silicio y el metal reaccionan para formar el siliciuro de manera semejante a la recombinación de intersticiales y vacantes

$$\frac{\partial C}{\partial t} = \nabla\left(D \cdot \nabla C\right) - R \quad . \tag{16.18}$$

donde C es la concentración de silicio/metal, R es la recombinación en volumen de moléculas de silicio/metal y D es el coeficiente de difusión del silicio/metal en el siliciuro.

La formación del siliciuro trae como consecuencia una reducción drástica del volumen que puede originar tensión en el siliciuro. El cambio de volumen asociado a esta reacción se modela utilizando los volúmenes atómicos de las especies que reaccionan y el volumen molecular del producto

$$xM + ySi \rightarrow M_x Si_y \quad . \tag{16.19}$$

El cambio de volumen viene dado por:

$$\Delta V = \frac{\left(xV_M + yV_{Si}\right) - V_{M_x Si_y}}{xV_M + yV_{Si}} \times 100 \tag{16.20}$$

donde ΔV es el cambio de volumen molecular, $\boldsymbol{V_M}$, $\boldsymbol{V_{Si}}$, $V_{M_x Si_y}$ el volumen molecular del metal, el silicio y el siliciuro respectivamente y (x, y) o las proporciones estequiométricas, el número de átomos de metal y silicio en el siliciuro $M_x Si_y$.

16.5 Modelos de Implantación Iónica

Los simuladores comerciales de procesos utilizan técnicas analíticas y estadísticas para modelar la implantación de iones. Por defecto, se utilizan los modelos analíticos basados en la reconstrucción de perfiles implantados a partir de la distribución de momentos calculados o medidos. La técnica estadística utiliza el cálculo basado en el método Monte Carlo de la trayectoria del ión para averiguar la distribución final de partículas.

Una distribución bidimensional se puede considerar como la convolución de una distribución longitudinal unidimensional (a lo largo de la dirección del implante) y de otra distribución transversal unidimensional (perpendicular a la dirección del implante). A continuación se describen tres modelos de implantación unidimensional, dos modelos de distribución transversal y un método para el cálculo del perfil implantado bidimensional.

Gaussiano

La manera más sencilla de construir un perfil unidimensional es utilizando una distribución Gaussiana:

$$C(x) = \frac{\phi}{\sqrt{2\pi}\Delta R_p} \exp \frac{-\left(x - R_p\right)^2}{2\Delta R_p^2} \tag{16.21}$$

donde ϕ es la dosis de iones por centímetro cuadrado, R_p es el rango proyectado y ΔR_p es la desviación estandar.

Pearson

Normalmente, la distribución gaussiana no es apropiada porque en la mayoría de los casos los perfiles reales son asimétricos. El método más simple y más utilizado para el cálculo de perfiles asimétricos producidos mediante implantación iónica es la distribución de Pearson, en particular la función de Pearson IV. El simulador Athena utiliza esta función para obtener perfiles longitudinales de la implantación. La función de Pearson hace referencia a una familia de curvas de distribución obtenidas al resolver la siguiente ecuación diferencial

$$\frac{df(x)}{dx} = \frac{(x-a)f(x)}{b_0 + b_1 x + b_2 x^2} \tag{16.22}$$

donde $f(x)$ es la función de la frecuencia. Las constantes a, b_0, b_1 y b_2 están relacionadas con los momentos de la función $f(x)$.

Dual Pearson

Para extender la aplicabilidad del modelo analítico hacia aquellos perfiles afectados por el fenómeno denominado channeling, se ha propuesto el método dual (o doble) de Pearson. Con esta expresión, la concentración de impurezas implantadas se calcula como una combinación lineal de dos funciones de Pearson

$$C(x) = \phi_1 f_1(x) + \phi_2 f_2(x) \tag{16.23}$$

donde la dosis es representada por cada función de Pearson $f_{1,2}(x)$. $f_1(x)$ y $f_2(x)$ están normalizadas, cada una con su propio conjunto de momentos. La primera función de Pearson representa la parte de dispersión aleatoria alrededor del pico del perfil y la segunda función representa la región de la cola del canal.

Convolution Method

Athena calcula perfiles implantados bidimensionales usando un método de convolución. La dirección de la implantación dentro del plano de simulación queda descrito mediante los ángulos de inclinación θ y rotación φ. θ es el ángulo entre la dirección del haz de iones y el eje *Y*, φ es el ángulo entre la dirección del haz de iones y el plano de simulación.

La manera más sencilla de construir una distribución bidimensional es mediante el producto de una función longitudinal $f_l(x)$ que puede ser una Gaussiana, Pearson o doble Pearson y una función transversal $f_t(y)$ independiente de la profundidad

$$f_{2D}(x,y) = f_1(x) f_t(y) \ . \qquad (16.24)$$

La función $f_t(y)$ debe ser simétrica y presentar forma de campana. No obstante, en general la función transversal $f_t(y)$ no es independiente de la profundidad porque existe una correlación muy fuerte entre el movimiento longitudinal y transversal de los iones implantados. Es posible tener en cuenta esta correlación utilizando una función transversal con desviación estándar lateral dependiente de la profundidad $\sigma_y(x)$. La mejor aproximación para $\sigma_y(x)$ es la función parabólica

$$\sigma_y^2(x) = c_o + c_1(x - R_p) + c_2(x - R_p)^2 \ . \qquad (16.25)$$

Los resultados de simulaciones Monte Carlo han demostrado que en la mayoría de los casos la función de distribución transversal f_t no es Gaussiana. Para poder reproducir estos resultados se ha trabajado con varias distribuciones no Gaussianas. Se ha encontrado que las funciones Pearson simétricas reproducen satisfactoriamente los resultados de las simulaciones Monte Carlo y presentan ventajas en el cálculo numérico. Otra buena alternativa es la función Gaussiana modificada (MGF). La elección de una función u otra no está clara y normalmente se realiza mediante comparaciones con distribuciones obtenidas a partir de simulaciones MC. Athena suele utilizar un promedio de las dos funciones anteriormente citadas.

Implantes Monte Carlo

Los modelos analíticos descritos en la sección anterior dan resultados muy buenos cuando se utilizan en implantaciones iónicas sobre estructuras planas (silicio desnudo o cubierto por una delgada lámina de otro material). Sin embargo, en estructuras formadas por muchas capas de diferentes materiales, geometrías que no sean planas y en los casos que no se han estudiado todavía experimentalmente se requieren modelos más sofisticados. La aproximación más flexible y extendida para simular la implantación de iones en condiciones no estándar es la técnica Monte Carlo. Este método permite el cálculo de los perfiles de implantación en una estructura arbitraria con exactitud comparable a la de los modelos analíticos en estructuras de una capa. Athena utiliza dos modelos para la simulación Monte Carlo de la implantación iónica: Materiales amorfos y materiales cristalinos. Los dos se basan en la aproximación binaria de la colisión (BCA), pero aplican distintas aproximaciones a la estructura del material y a la propagación del ión en ella.

1.- Naturaleza física del problema

Un haz de iones rápidos (en un rango de energías aproximado de entre 50 eV/amu y 100 keV/amu) penetrando en el sólido cristalino o amorfo es frenado y dispersado por las colisiones nucleares y la interacción electrónica. A lo largo de su trayectoria, un proyectil individual interacciona con los átomos del material objetivo que a su vez

pueden iniciar una cascada de colisiones. Éstos pueden abandonar la superficie (sputtering) o depositarse en un sitio diferente del original. Junto con los proyectiles que se depositan en el substrato, estos procesos dan lugar a cambios locales de la composición, daños en la estructura cristalina y finalmente la amorfización del blanco. Dependiendo de la orientación cristalina, la dirección del haz y el tipo de proyectil implantado, los daños creados por ellos tienen diferente distribución espacial. Con flujos aún mayores, estos fenómenos provocarían cambios de la composición superficial y el establecimiento de un perfil fijo de los iones implantados.

2.- Método de Solución

Las trayectorias de las partículas móviles individuales y sus colisiones se modelan por medio de la aproximación binaria de la colisión para materiales cristalinos, policristalinos y amorfos, usando un potencial colombiano apantallado para las colisiones nucleares y una combinación de la aproximación local y no local del gas de electrones libres para la pérdida de energía electrónica. En cada colisión nuclear, se determina el parámetro de impacto y el ángulo azimutal de desviación según la estructura cristalina usando su simetría de translación. Para materiales amorfos, estos dos parámetros se determinan mediante la elección de números al azar. Se elige un escalado apropiado de modo que cada proyectil incidente (pseudo-proyectil) represente una fracción del flujo total implantado. Al finalizar la simulación de cada pseudo-proyectil y de de las colisiones en cascada asociadas, se calculan las concentraciones locales de la especie implantada, de las vacantes e intersticiales creadas según la matriz densidad correspondiente.

3.- Frenado electrónico

La interacción inelástica con el gas de electrones utilizado en la simulación consiste de dos mecanismos independientes, local y no local. Estos dos mecanismos de frenado electrónico son absolutamente diferentes en naturaleza y comportamiento ya que presentan una dependencia energética y espacial distinta. Las pérdidas de energía inelásticas locales se basan en el modelo propuesto por Firsov. En este modelo la estimación de la energía electrónica perdida por la colisión se basa en la suposición de una imagen cuasi-clásica de los electrones (la energía media de excitación de las capas electrónicas y la distribución y movimientos de los electrones siguen el modelo de Thomas-Fermi del átomo). En este modelo cuasi-clásico, la transferencia de energía del ión al átomo es debida al paso de electrones de una partícula a la otra, resultando en un cambio de momento del ión.

Monte Carlo para materiales amorfos

En el dopado de semiconductores la distribución en reposo de las impurezas implantadas es de enorme importancia. La manera más sencilla de describir la penetración de iones en objetivos amorfos se realiza usando un modelo estadístico del transporte. Monte Carlo es el

procedimiento más conveniente cuando se utilizan múltiples componentes (tanto de impurezas como de objetivos) y pretendemos obtener una distribución final en dos o tres dimensiones. La distribución resultante se construye después de simular un gran número de trayectorias ya que la precisión estadística es proporcional a $\sqrt{N}$. Mientras que el ión penetra un sólido, experimenta una secuencia de colisiones con los átomos del objetivo hasta que alcanza el reposo. Un modelo simplificado de estas interacciones es una secuencia de colisiones nucleares binarias instantáneas separadas por una cierta distancia en línea recta (longitud del camino libre de vuelo) durante la cual el ión experimenta una pérdida continua de energía electrónica (no-local). Las colisiones se consideran independientes, es decir, el estado de un ión después de una colisión dependa únicamente del estado del ión antes de la colisión. El modelo supone que la distribución de los átomos del blanco se selecciona al azar después de cada colisión (es decir, el blanco no tiene ninguna estructura y ninguna memoria). Consecuentemente, una secuencia de colisiones se describe seleccionando aleatoriamente la localización de la siguiente colisión dependiendo de la localización y velocidad del ión.

Monte Carlo para materiales cristalinos

Para calcular la distribución en reposo de los proyectiles, Athena simula colisiones atómicas en objetivos cristalinos usando la aproximación binaria de la colisión (BCA). El algoritmo sigue la secuencia de los iones lanzados desde un haz externo sobre el material objetivo. Los objetivos pueden tener regiones compuestas de diferentes materiales, cada uno con su propia estructura cristalina o amorfa. El frenado de los proyectiles se sigue hasta que abandonan el material o su energía disminuye por debajo de un cierto valor umbral.

Daños producidos por implantación iónica

Los daños sobre la estructura cristalina provocados por la implantación de iones pueden desempeñar un papel importante en los diferentes mecanismos relacionados con la difusión y la oxidación. Athena presenta varias maneras de incluir los daños generados para que se puedan utilizar en un cálculo posterior de la difusión. Los daños inducidos por la implantación tienen su origen en las colisiones atómicas en cascada. Si estas colisiones en cascada alcanzan elevadas densidades puede dar lugar a una transformación del material cristalino en otro amorfo. La simulación precisa de la colisión en cascada junto con una estimación simultánea de la generación de diferentes tipos de defectos puntuales, clusters y de defectos espaciales se puede realizar solamente en la aproximación de colisión binaria (BCA) o con simuladores de dinámica molecular (MD). Estas simulaciones consumen gran cantidad de tiempo y recursos y su uso no es práctico dentro de los simuladores de propósito general. Generalmente, los daños generados y la distribución de los

defectos dependen de la energía, de la especie y de la dosis de iones implantados.

16.6 Otros modelos

Modelos de grabado

SSuprem4 considerada el grabado como un problema puramente geométrico. El ataque químico se simula como un proceso a baja temperatura donde no se produce redistribución de impurezas. Se debe indicar el material y los contornos geométricos de la región a grabar y existen diferentes formas de hacerlo:

- Indicando las coordenadas (*x,y*) de los vértices del polígono que contiene la región de interés.
- Se puede indicar una región a la derecha o a la izquierda de una línea recta hasta el límite de la estructura. Para ello se deben especificar las coordenadas de ese segmento.
- También es posible realizar un grabado sobre todas las regiones que contengan un material particular o sobre ese material en una región determinada.

Simulación en SiGe

En la mayoría de los casos, las capas de $Si_{1-x}Ge_x$ con un contenido gradual o constante de germanio se forman mediante un proceso especial de epitaxia. Varios experimentos revelan que la difusión del boro en aleaciones de $Si_{1-x}Ge_x$ es diferente de la difusión en un substrato de silicio puro. Estos experimentos también demostraron que el coeficiente de difusión del boro depende de la concentración de germanio, *x*. Para simular estos efectos se utiliza un modelo especial para la difusión del boro en silicio.

El modelo empírico tiene en cuenta dos hechos experimentalmente demostrados: La difusividad del B disminuye al aumentar el contenido de germanio y la concentración intrínseca de electrones n_i aumenta con el contenido de germanio (menor E_g del Ge). Estos dos factores se incluyen en los simuladores numéricos mediante las siguientes expresiones:

$$D_{BI}^{x} = DIX.0 \cdot \exp\left(-\frac{DIX.E + x \cdot EAFACT \cdot SIGE}{kT}\right) \qquad (16.26)$$

donde el coeficiente de difusión asociado a la pareja intersticial-boro disminuye exponencialmente con el contenido de germanio (*x*).

$$n_i\left[Si_{1-x}Ge_x\right] = n_i\left[Si\right]\left(1 + x \cdot NIFACT \cdot SIGE\right) \qquad (16.27)$$

donde la concentración intrínseca de electrones en $Si_{1-x}Ge_x$ aumenta linealmente con el contenido de germanio.

RESUMEN

En este capítulo se han estudiado los modelos que se utilizan en la simulación de procesos. Para ello se parte de todo lo estudiado previamente en los capítulos dedicados a la tecnología de fabricación de dispositivos electrónicos donde se analizan con detalle los distintos modelos matemáticos utilizados para explicar cada uno de los procesos físicos/químicos implicados (oxidación, difusión, implantación iónica, grabado, etc). La resolución numérica de las ecuaciones implica una limitación importante ya que se observa que los modelos más precisos implican una gran demanda de recursos computacionales así como de tiempo de cálculo. Por este motivo es habitual encontrar diferentes posibilidades para un mismo proceso dependiendo de la precisión que se necesita y de los recursos disponibles.

REFERENCIAS

[1] S. Wolf. *Silicon Processing for the VLSI Era, Volume 4.* Lattice Press, 2002.

[2] *Athena User's Manual 2D Process Simulation Software*. Silvaco International, Santa Clara, CA USA December 2002.

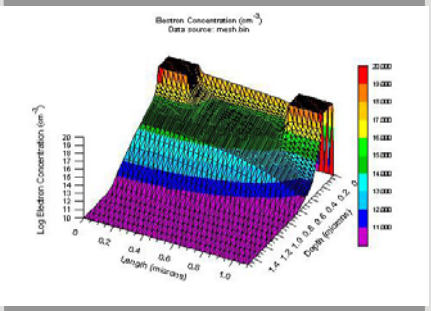
[3] http://www.silvaco.com

17

Capítulo

Simulación de Dispositivos con PISCES

MOSFET (POSTMINI)



Índice

Objetivos

- Presentar el simulador de dispositivos PISCES
- Describir los modelos físicos incluidos
- Comentar los distintos modelos de movilidad y Generación-Recombinación incluidos.
- Enumerar los distintos tipos de simulación disponibles.
- Describir la estructura de los archivos de entrada que caracterizan los dispositivos a simular.
- Introducir la herramienta POSTMINI para la posterior presentación de los resultados de la simulación.
- Desarrollar algunos ejemplos de simulaciones en dispositivos anteriormente estudiados en capítulos anteriores.

Palabras Clave

Simulador PISCES
Modelos de movilidad
Modelo de Difusión-Deriva
Modelo DUET
Variables de estado para el modelo DUET
Modelos de movilidad para bajos campos
Modelos de movilidad dependientes del campo transversal
Modelos de movilidad dependientes del campo longitudinal
Modelos de Generación-Recombinación
Simulación estacionaria
Simulación AC
Simulación transitoria
Archivo de entrada
Herramienta POSTMINI
Representación 1D, 2D, 3D

17.1 Desarrollo de PISCES

PISCES constituye una de las herramientas más utilizadas tanto en la industria como en investigación para la simulación de dispositivos semiconductores. Existen diferentes versiones disponibles tanto *freeware* como de pago constituyendo la base de algunos de los módulos de programas comerciales tan usados como el ATLAS comercializado por Silvaco.

La última versión de PISCES, desarrollada por un grupo de la Universidad de Stanford (USA), implementa modelos físicos avanzados para dispositivos que han ido apareciendo desde la primera versión y métodos numéricos mejorados para aplicaciones 2D que permiten el desarrollo de las nuevas aplicaciones que están siendo investigadas tanto en la industria como en el entorno universitario.

Básicamente este desarrollo se centra en tres objetivos principales:

- Desarrollar un nuevo código para PISCES.
- Mejorar los modelos de movilidad para estructuras MOS.
- Desarrollo de un modo mixto de simulación.

Como resultado se ha creado una nueva versión denominada PISCES-2ET debido al modelo que implementa, Dual Energy Transport (DUET o 2ET), que será comentado más adelante.

Partiendo de las versiones de PISCES II 9009 de Stanford y de la 8830 de Intel, en PISCES-2ET se han incluido algunas nuevas capacidades ausentes en versiones anteriores que añaden más flexibilidad tanto en las estructuras a simular como en los modelos utilizables. Entre las más importantes se encuentran:

- Un completo modelo de transporte de energía de portadores válido para semiconductores simples y materiales compuestos. En combinación con el modelo de difusión térmica para la red cristalina es posible predecir fenómenos físicos tales como el de sobredisparo en la velocidad y efectos de portadores calientes que son críticos en la determinación de las características eléctricas de dispositivos submicra.
- Capacidad de simulaciones completas para dispositivos basados en heteroestructuras. Simulaciones DC, AC y transitorias para dispositivos multicapa y formados por distintos materiales. En el simulador es posible encontrar los parámetros correspondientes a materiales tales como SiGe, AlGaAs, AlInAs, y GaInAsP.
- Análisis AC para altas frecuencias utilizando el algoritmo iterativo TFQMR pudiéndose estudiar las características del dispositivo en frecuencias cercanas a las de corte sin problemas adicionales de memoria o convergencia.

- Han sido incluidos modelos de movilidad dependientes de la energía de los portadores y modelos de ionización por impacto así como modelos semiempíricos de movilidad dependientes del campo eléctrico tanto transversal como longitudinal.
- Finalmente, se ha mejorado la solución inicial para el Método de Newton usando el denominado esquema proyectivo de Newton. Este esquema también es utilizado en el análisis AC a frecuencia cero fundamental para el análisis mixto de circuitos/dispositivos.

El modo mixto de simulación es una de las más importantes aportaciones nuevas que incluye esta versión. Gracias a él es posible generar los modelos de SPICE de los dispositivos simulados con PISCES de forma que se puede predecir su comportamiento en circuitos.

Todos los comentarios que aparecen en este capítulo están realizados sobre la base del PISCES-2ET. Sin embargo las simulaciones y el código disponible para evaluación corresponden a la versión PISCES-IIB. Por tanto, todas las características simuladas en los ejemplos son compatibles con PISCES-2ET y IIB aunque no todas las posibilidades descritas para PISCES-2ET están disponibles en la versión IIB que solo incluye el modelo de Difusión-Deriva.

17.2 Descripción del Simulador

PISCES es un simulador para dispositivos semiconductores capaz de resolver el transporte en estructuras tanto uni y bidimensionales. El modelo utilizado en las primeras versiones fue el de difusión-deriva (DD) comentado en el Capítulo 16.

En PISCES-2ET se utiliza el modelo DUET basado en la aproximación de momentos para resolver la ecuación de transporte de Boltzmann (BTE). Para describir el dispositivo semiconductor se utilizan seis variables de estado:

- Potencial electrostático, ψ
- Concentraciones de portadores, n y p
- Temperatura de los portadores, T_n y T_p
- Temperatura de la red cristalina, T_L

Todas ellas son funciones del espacio y del tiempo y se consideran como variables independientes. El resto de magnitudes tales como la corriente son calculadas una vez conocidas las seis anteriores.

Para conocer las distribuciones de las distintas variables bajo unas condiciones de polarización dadas es necesario resolver seis ecuaciones acopladas con las condiciones de contorno apropiadas.

Con el modelo de difusión y deriva (DD) para el transporte de portadores y usando las ecuaciones de continuidad y la de Poisson es

posible obtener las distribuciones de potencial electrostático, ψ, y las de portadores, *n* y *p*.

Para el caso de la temperatura se introducen tres nuevas ecuaciones derivadas de las ecuaciones de balance de energía que dan cuenta de los procesos de difusión de la temperatura tanto de los portadores como de la red cristalina.

Modelos de Movilidad

La movilidad de los portadores es uno de los parámetros más importantes para el modelado del transporte. En PISCES-2ET es modelada principalmente como función de la densidad total de dopantes, la temperatura de la red cristalina, mecanismos de dispersión superficial y en interfases dependientes de los campos transversales y finalmente, la dependencia con el campo longitudinal.

En el caso de dispositivos submicra la dependencia con el campo longitudinal puede no ser suficientemente precisa y deben añadirse efectos no locales que, principalmente, están caracterizados por la dependencia con la temperatura de los portadores. En general la dependencia de la movilidad puede expresarse de la siguiente forma:

$$\mu\left(N, T_L, E_\perp, E_\parallel / T_c\right) = f\left(\mu_0\left(N, T_L, E_\perp\right), E_\parallel / T_c\right) \qquad (17.1)$$

Donde $E_\perp$ y $E_\parallel$ corresponde con la componente transversal y longitudinal del campo eléctrico con respecto a la dirección de la corriente. *N* es el dopado, μ_0 es denominada movilidad a bajo campo porque cuando $E_\parallel \to 0, \mu \to \mu_0$. T_c representa la temperatura de los portadores y el símbolo $E_\parallel / T_c$ indica que la dependencia se da con uno de los dos factores pero nunca con los dos a la vez.

A continuación se presentan los diferentes modelos implementados en PISCES-2ET.

Modelos de Baja Movilidad:

- Movilidad Constante: En él se utiliza una movilidad a bajo campo constante que sólo depende del material que se esté utilizando.
- Modelo dependiente del dopado: Cuando se active, la movilidad se calcula a partir una tabla de valores dependiente del dopado. En caso de que el valor no aparezca se realiza una interpolación. Si el material no es silicio se aplica el modelo de movilidad constante.
- Modelo analítico dependiente del dopado: En este modelo se tienen en cuenta dos términos. El primero dependiente de la movilidad superficial y el segundo de la movilidad en volumen. El peso de cada uno de los términos depende de la distancia a la interfase Si-SiO_2.

- Modelo de Arora: Modelo empírico basado en el ajuste de curvas de movilidad a diferentes temperaturas de la red cristalina. Existe la posibilidad de personalizar el modelo utilizando unos parámetros distintos de los propuestos en el modelo original.
- Modelo de dispersión "carrier-carrier": Este modelo sigue la aproximación de Dorkel y Leturq cuya formulación incluye una dependencia de la movilidad con la concentración de los dos tipos de portadores.

Modelos dependientes del Campo Transversal

- Modelo de Campo Local de Intel: Este modelo existente en dos versiones incluye el modelo analítico para bajo campo más una corrección por campo perpendicular (representada por r_{perp}) dependiente de un campo crítico de la forma

$$r_{perp} = \left(1 + \frac{E_{\perp}}{E_{crit}}\right)^{-\beta} \qquad (17.2)$$

- Modelo de Lombardi: Obtenido a partir de un gran número de medidas experimentales realizadas por Lombardi y que da lugar a un modelo dependiente de tres factores. El primero es la movilidad a bajo campo calculada con cualquiera de los modelos anteriores (μ_0), el segundo tiene en cuenta los efectos de dispersión por fonones acústicos (μ_{ac}) y el tercero efectos debidos a la dispersión producida en la superficie e interfases (μ_{srf}). La movilidad final es obtenida aplicando la regla de Mathiessen.

$$\frac{1}{\mu(N, T_L, E_{\perp})} = \frac{1}{\mu_{ac}(N, T_L, E_{\perp})} + \frac{1}{\mu_{srf}(E_{\perp})} + \frac{1}{\mu_0(N, T_L)} \qquad (17.3)$$

- Modelo de Movilidad Superficial de Watt: Este modelo supone que toda la lámina en inversión se encuentra dentro del espacio situado entre la interfase con el óxido y el primer punto del mallado correspondiente al Si. Por tanto, al contrario de lo que se suele utilizar, el primer punto del mallado debe estar razonablemente separado de la interfase (unos cientos de Ángstrom) en vez de ser una zona densa desde el punto de vista de la discretización. La movilidad utilizada depende del campo efectivo que corresponde con el campo medio en la capa de inversión y es calculada de nuevo a partir de la regla de Mathiessen a partir de términos de dispersión por fonones, rugosidad superficial y dispersión por impurezas ionizadas en la lámina de inversión.

- Modelo de Campo Transversal no Local de Shin: Este modelo tiene en cuenta no sólo efectos locales, sino también otras propiedades no locales como el espesor de la lámina de inversión.

Modelos dependientes del Campo Longitudinal

Cuando los portadores están sometidos a campos longitudinales, la movilidad en general se ve afectada. Este cambio puede ser modelado como una función del campo local para el caso de que el campo no sea excesivamente elevado y dependiente de la temperatura de los portadores cuando el tiempo de relajación para el momento sea menor que para la energía.

- Modelo General para Si: Para el caso del Si se utiliza un modelo que corrige los efectos producidos por el campo transversal, de forma que a la movilidad a bajo campo (ya dependa esta del campo transversal o no) se le añade un término dependiente del campo longitudinal de la forma

$$\mu\left(N,T_L,E_\perp\right)=\mu_0\left(N,T_L,E_\perp\right)\left[1+\left(\frac{\mu_0\left(N,T_L,E_\perp\right)E_\square}{v_{sat}}\right)^{\beta}\right]^{-1/\beta} \qquad (17.4)$$

- Modelos para GaAs: Como se puede comprobar en la expresión (17.4) la velocidad de arrastre $\mu E_\square$ aumenta monótonamente con el campo longitudinal hasta que se satura. Sin embargo para el caso del GaAs y debido a la existencia de varios valles con diferentes masas efectivas la velocidad de los portadores va aumentando hasta alcanzar un pico y luego decrece aún cuando el campo longitudinal sigue creciendo.

Modelos dependientes de la Temperatura de los Portadores

- Modelo de Hänsch para Si: Como se ha comentado con anterioridad existe la posibilidad de utilizar una dependencia con la temperatura de los portadores en lugar de con el campo local longitudinal. Uno de los modelos, propuestos por Hänsch, viene dado por la siguiente expresión

$$\mu\left(N,T_L,E_\perp,T_c\right)=\frac{\mu_0\left(N,T_L,E_\perp\right)}{1+\frac{3}{2}\alpha\left(N,T_L,E_\perp\right)k_B\left(T_c-T_L\right)} \qquad (17.5)$$

Modelos de Generación-Recombinación

- Modelo de Ionización por Impacto: Este modelo da cuenta de los procesos de generación de pares electrón-hueco debidos al impacto de portadores contra átomos

de la red siendo su expresión general de la forma siguiente:

$$G = \alpha_n n v_n + \alpha_p p v_p \qquad (17.6)$$

En PISCES existen diferentes modelos para calcular los valores de α teniéndose en cuenta dependencias con el campo local o la temperatura de los portadores.

- Modelo SRH: El modelo Shockley-Read-Hall se ha incluido en PISCES con la expresión ya comentada en el Capítulo 16.

$$U_{SRH} = \frac{np - n_i^2}{\tau_p \left[n + n_i \, e^{\left(\frac{q(E_t - E_i)}{kT} \right)} \right] + \tau_n \left[p + p_i \, e^{\left(\frac{q(E_i - E_t)}{kT} \right)} \right]} \qquad (17.7)$$

- Modelo de Recombinación Auger: Este es un proceso de recombinación en el que intervienen tres partículas, dos electrones y un hueco o dos huecos y un electrón.
- Foto-Generación: En este caso se modela un término de generación forzada por el flujo de fotones que inciden sobre el semiconductor. La tasa de generación es proporcional a dicho flujo y decae exponencialmente con la profundidad en el dispositivo a través de la trayectoria de incidencia. Este decaimiento está caracterizado por el coeficiente de absorción.

17.3 Tipos de Simulación

PISCES puede realizar tres tipos de simulaciones: DC, AC y transitoria. La descripción de los parámetros para cada uno de los tipos viene dado por el comando SOLVE. Su sintaxis es la siguiente:

SOlve <estimate> <dc bias> <transient> <ac> <files>

<estimate>

INitial = <logical>: El primer punto de polarización para una estructura dada debe tener este parámetro especificado.

PREvious = <logical>: Utiliza como solución inicial la anterior conocida en el caso de que se esté variando la polarización de uno de los terminales.

PROject = <logical>: Realiza una extrapolación a partir de dos soluciones precedentes de forma que se obtenga una solución inicial mejorada. Después de dar un polarización inicial PISCES II usa esta extrapolación donde sea posible.

LOcal = <logical>: Si no hay parámetro de estimación de la solución inicial, es posible utilizar una estimación basada en los valores locales de los psudoniveles de Fermi.

<bias conditions> Para la polarización se pueden utilizar los siguientes valores:

V1 … V0 = <real> Polarización para contactos controlados por tensión.

I1 … I0 = <real> Corriente (en A/m) para terminales controlados por coriente.

VStep/ IStep = <real> (por defecto 0.0) Representa incrementos en la tensión (corriente) al terminal o terminales especificados por el identificador entero de la etiqueta de éstos asignado en ELECTRODE.

NSteps = <integer> (por defecto 0) Representa el número de veces en que se va a incrementar la tensión/corriente en un electrodo dado.

Electrode = <integer> Es un número entero de n dígitos donde cada uno de ellos corresponde con un número de electrodo. (Nota: si hay 10 electrodos no se debe poner el número 0 en primer lugar).

N.bias/ P.bias = <real> Especifica los pseudoniveles de Fermi para los portadores que no sean resueltos. Si este parámetro no es especificado los valores son asignados en función del dopado.

<transient>

Dt or TSTEp = <real> (por defecto 0) Especifica el incremento de tiempo que se utilizará.

TSTOp o TFinal = <real> Representa el tiempo para el que la simulación se da por finalizada. Alternativamente se puede utilizar NSTEPSde forma que el tiempo final de simulación viene dado por

$$t_{final} = t_0 + NSTEPS \times TSTEP \tag{17.8}$$

Ramptime/ENdramp = <real> (por defecto 0) RAMPTIME y ENDRAMP aplican variaciones lineales para cualquier cambio en la polarización. RAMPTIME indica la duración de la transición mientras que ENDRAMP indica el instante temporal en el que ésta termina.

<ac>

AC.analysis = <logical> (por defecto false). Flag que indica el análisis en pequeña señal que se realiza después de que se resuelva la condición inicial en DC.

FRequency = <real>. Indica a la frecuencia a la que se realiza el análisis.

FStep = <real> (por defecto 0). El análisis puede ser repetido a diferentes frecuencias sin tener que resolver la condición inicial en DC de nuevo cuando se especifica el paso de frecuencias. A la frecuencia

resuelta en el paso anterior se le añade el valor de FSTEP o se multiplica por este factor si está activado MULT.FREQ.

MUlt.freq = <logical> (por defecto false)

NFsteps = <integer> (por defecto 0) Indica el número de veces que se realiza el incremento dado por FSTEP.

VSs = <real> (por defecto 0.1*kT/q) Es la amplitud de la pequeña señal aplicada.

TErminal = <integer> (por defecto todos) Es el contacto al que se aplica la señal AC. Es posible especificar más de un contacto utilizando la concatenación aunque cada uno de ellos será resuelto por separado.

S.omega = <real> (por defecto 1.0). El método de resolución utilizado es denominado SOR (Successive Over Relaxations) donde esta variable da el valor de la constante de relajación.

MAx.inner = <integer> (por defecto 25) corresponde con el número máximo de iteraciones realizadas con SOR.

TOlerance = <real> (por defecto 1 x 10^{-5}) indica el criterio de convergencia para la iteración SOR.

<files> (optional)

Outfile = <filename>. Con este parámetro se especifica el nombre del archivo binario de salida en donde se almacena la solución del punto de polarización. El nombre puede contener hasta 20 caracteres. Si se resuelve para más de un punto de polarización se añade un carácter al final indicando el paso al que corresponde la solución.

Currents = <logical> (por defecto false). Si se activa esta opción, se calculan las corrientes de electrones, huecos y de desplazamiento así como el campo eléctrico guardándose en el archivo de solución.

AScii = <logical> (por defecto false). Cuando esta opción está activada el archivo de salida resulta ser de tipo ASCII en lugar de binario.

17.4 El archivo de entrada

En PISCES las características del dispositivo a simular, los modelos a utilizar y las simulaciones a realizar son especificados en un archivo de entrada generado por el usuario.

El formato utilizado debe seguir ciertas pautas de forma que pueda ser leído por GENII que es el procesador de entradas creado por Stanford y que es utilizado también para SUPREM.

Cada línea constituye una orden específica y es identificada mediante la primera palabra. El resto de la línea son los diferentes parámetros correspondientes al comando introducido con anterioridad. Las palabras en cada línea son separadas por espacios en blanco o tabuladores. Si en cualquier momento fuese necesario utilizar más de una línea se debe utilizar un signo más (+) en el primer espacio no blanco de la nueva línea.

No es necesario escribir el nombre completo de los caracteres, basta un número suficiente para que sean identificados de forma unívoca.

Existen tres tipos de parámetros: numéricos, lógicos y caracteres. Los parámetros numéricos son asignados colocando junto al nombre un signo igual (=) y a continuación el valor numérico deseado. En el caso de los caracteres, después del signo igual se acompaña la cadena de caracteres que se considera terminada en cuanto se encuentre un espacio blanco o un tabulador. Los valores lógicos se determinan de la siguiente manera; si aparece el nombre del parámetro se considera que el valor de la variable es TRUE, si éste va precedido por un asterisco (*) se considera que está negado y, por tanto, toma el valor FALSE.

En algunos casos, el orden de aparición de los distintos comandos es importante, debido a que se dan ciertas dependencias entre unos y otros. Las diferentes reglas que deben ser respetadas son las siguientes:

- MESH debe preceder al resto de los comandos salvo TITLE, COMMENT y OPTIONS.
- Cuando se define un mallado rectangular se debe realizar en el siguiente orden:
  - ✓ MESH
  - ✓ X.MESH (todos los puntos)
  - ✓ Y.MESH (todos los puntos)
  - ✓ ELIMINATE
  - ✓ SPREAD
  - ✓ REGION
  - ✓ ELECTRODE
- ELIMINATE y SPREAD son opcionales, pero si aparecen deben hacerlo en ese orden.
- DOPING debe aparecer justo después de la definición del mallado.
- Antes de calcular una solución es necesario realizar una factorización simbólica. A no ser que se calcule una solución para el caso en equilibrio se debe utilizar una previa para aportar una estimación inicial.
- Cualquier CONTACT debe preceder a la orden SYMBOLIC.
- Los parámetros físicos no deben ser cambiados utilizando MATERIAL, CONTACT o MODEL después de haber utilizado una vez SOLVE o LOAD. MATERIAL y CONTACT deben preceder al comando MODEL.
- PLOT 2D debe preceder a cualquier trazado de contornos para establecer los límites del trazado.
- PLOT 2D, PLOT 1D, REGRID o EXTRACT necesitan acceder a las variables de resolución (potencial

electrostático, concentraciones y temperaturas), por tanto deben ser precedidos por LOAD o SOLVE de forma que estas cantidades estén disponibles.

17.5 La Herramienta POSTMINI

POSTMINI es una herramienta gráfica de postprocesado para simuladores de dispositivos y procesos. Con ella es posible leer los archivos guardados con diferentes simuladores tales como MINIMOS, MEDICI, TSUPREM4 o PISCES y permite al usuario visualizar las diferentes cantidades almacenadas en los archivos de salida. Asimismo es posible importar datos desde archivos de tipo ASCII o calcular funciones de forma analítica sirviendo como un programa de propósito general para trazar curvas y superficies.

Los gráficos realizados están disponibles para estaciones de trabajo que trabajen con X11 y con dispositivos PostScript (tanto monocromo, color y de forma encapsulada). También es posible crear gráficos a partir de comandos escritos en un fichero.

A continuación se presentan algunas de las funciones disponibles en POSTMINI:

- 1D – Plot: Dibuja una representación X-Y de una serie de datos correspondientes a un fichero 1D o una sección de un archivo que represente datos 2D para cualquier línea vertical u horizontal.
- 2D – Plot: Representa un diagrama de contorno de un archivo de datos 2D.
- 3D – Plot: Representa una cantidad como una superficie en 3D.
- Compare: Dibuja varias curvas sobre el mismo gráfico pudiendo proceder éstas del mismo archivo o de diferentes. También es posible dibujar diagramas de barras.
- Overlay: Permite representar varios diagramas de contorno sobre la misma gráfica.
- Find: Busca el punto en que una determinada cantidad alcanza un valor.
- Integrate: Integra una determinada cantidad en una región o a lo largo de una línea.
- Line: Escribe en un archivo una sección de un conjunto de datos bidimensional a lo largo de cualquier línea vertical u horizontal.
- Minmax: Determina el valor máximo o mínimo de una cantidad interna.
- Print: Escribe un conjunto de datos bidimensionales en un formato determinado en un archivo.

- Read: Lee datos de un archivo. En primer lugar se pide el tipo de fichero que se va a leer (en nuestro caso PISCES). Después el archivo en el que se almacena el mallado (generalmente MESH.BIN) y por último el archivo con los datos que se van a representar.
- Show: Muestra información sobre la simulación que se ha realizado (actualmente sólo para archivos leídos desde MINIMOS).
- Default: Cambia los atributos por defecto de POSTMINI.
- Restore: Restaura un gráfico desde un archivo POSTMINI con diferentes comandos.
- Shell: Ejecuta comandos del sistema operativo sin salir de POSTMINI.
- Save: Archiva el gráfico en uso en un archivo de POSTMINI.
- Window: Gestiona múltiples gráficos en estaciones de trabajo.
- Exit, Quit: Finaliza POSTMINI.

17.6 Ejemplos

A continuación se presentan varios ejemplos de utilización de PISCES y de visualización de los resultados con POSTMINI. Es muy recomendable variar distintos parámetros de las estructuras aquí presentadas de forma que es posible ver cómo se modifican ciertas magnitudes al cambiar las condiciones de polarización o la geometría del dispositivo.

Ejemplo 17.1: Diodo Largo

Como ejemplo de iniciación se presenta el código para un diodo largo utilizando el método de Newton para su resolución y con el modelo SRH para generación-recombinación. El código para la simulación es el siguiente

```
COMMENT Diodo pisces
MESH RECTANG NX=35 NY=20 OUTF=MESH.BIN
X.MESH N=1 LOC=0
X.MESH N=35 LOC=1.0


Y.MESH N=1 LOC=0

Y.MESH N=20 LOC=0.1



COMMENT  DEFINICION zona n
REGION NUM=1 IX.LO=1 IX.HI=18 IY.LO=1 IY.HI=20 SILICON

COMMENT DEFINICION zona p
REGION NUM=2 IX.LO=18 IX.HI=35 IY.LO=1 IY.HI=20 SILICON


COMMENT ELECTRODO ZONA N
ELECTRODE NUM=1 IX.LO=1 IX.HI=1 IY.LO=1 IY.HI=20

COMMENT ELECTRODO ZONA P
ELECTRODE NUM=2 IX.LO=35 IX.HI=35 IY.LO=1 IY.HI=20


DOP REGION=1 UNIFORM CONC=1E16 N.TYPE
DOP REGION=2 UNIFORM CONC=1E16 P.TYPE

COMMENT MODELO DE G-R SRH

MODELS CONSRH

 MATERIAL  REGION=1  MUN=700  TAUN0=  1e16  TAUP0=1e16  NSRHN=1
NSRHP=1
 MATERIAL  REGION=2  MUN=700  TAUN0=  1e16  TAUP0=1e16  NSRHN=1
NSRHP=1
```

```
MATERIAL REGION=1 MUp=250
MATERIAL REGION=2 MUp=250

COMMENT METODO DE RESOLUCION DE NEWTON

SYMBOLIC NEWTON CARRIERS=1
METHOD IT=100

COMMENT ARCHIVO DE SALIDA DE CORRIENTE

LOG OUTF=CURRENT.DAT

SOLVE INIT

SOLVE V1=0  V2=0  OUT=SOL

SOLVE ELEC=2 VSTEP=0.1 NSTEP=20 out=0

SOLVE V1=0  V2=1  OUT=SOL1
SOLVE V1=0  V2=0.7 OUT=SOL07
```

Como se puede observar en el código se generan varios archivos de salida. El primero de ellos corresponde a la corriente para todos los valores de polarización simulados. Los que llevan por nombre SOL, SOL1 y SOL07 contienen información sobre potenciales y portadores para polarizaciones cátodo-ánodo de 0, 1 y 0.7 V respectivamente. Finalmente se realiza una simulación incrementando la diferencia de potencial en pasos de 0.1 V hasta 2V.

En la Figura 17.6.1 se puede observar la imagen obtenida tras procesar con POSTMINI los datos de la distribución de potencial para el dispositivo cuando no hay tensión aplicada en los terminales.

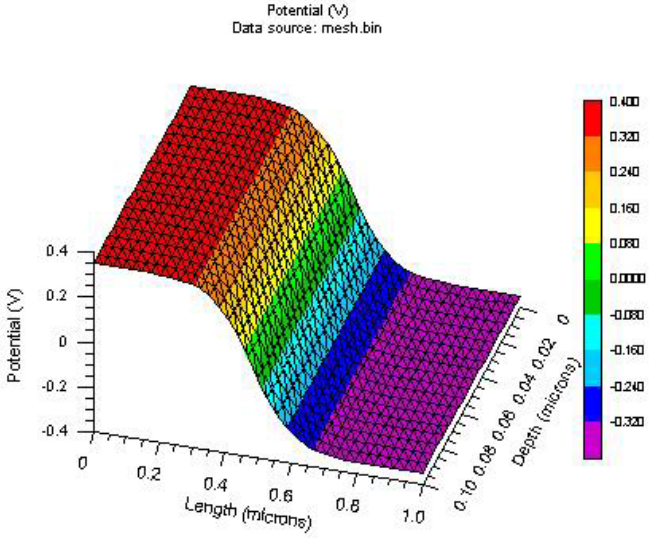


Figura 17.6.1 Potencial en el diodo sin polarizar

En la Figura 17.6.2 se muestra la relación I–V para el diodo con los datos obtenidos a partir del archivo CURRENT.DAT.

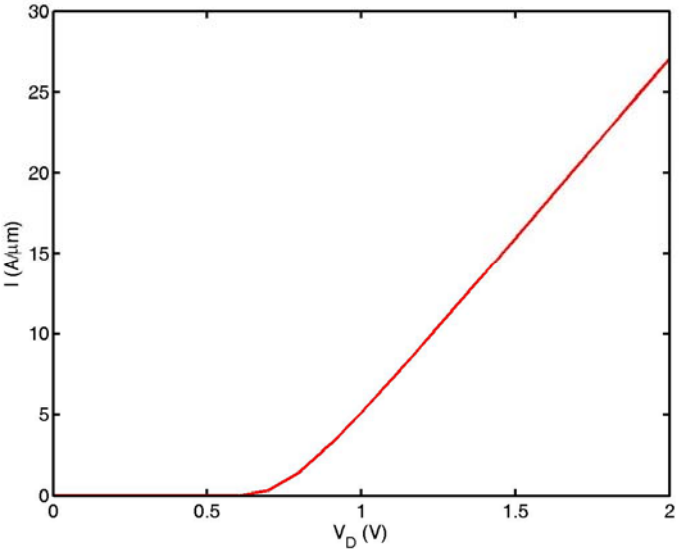


Figura 17.6.2 Característica I-V para el diodo simulado

Ejemplo 17.2. Transistor MOSFET

El siguiente ejemplo corresponde a una estructura MOSFET en la que se ha incluido no solo la zona de canal, sino también una parte del Si del substrato. El código para simular la estructura incluyendo una simulación transitoria para cambiar la tensión en la puerta es el siguiente:

```
COMMENT MOSFET pisces
MESH RECTANG NX=35 NY=45 OUTF=MESH.BIN
X.MESH N=1 LOC=0
X.MESH N=35 LOC=1.1


Y.MESH N=1 LOC=0
Y.MESH N=3 LOC=0.02
Y.MESH N=35 LOC=0.5
Y.MESH N=45 LOC=1.5

REGION NUM=1 IX.LO=1 IX.HI=35 IY.LO=1 IY.HI=3 OXIDE

COMMENT  DEFINICION DEL CANAL de arriba
REGION NUM=2 IX.LO=7 IX.HI=28 IY.LO=3 IY.HI=15 SILICON

COMMENT DEFINICION CANAL DE ABAJO
REGION NUM=3 IX.LO=1 IX.HI=35 IY.LO=15 IY.HI=45 SILICON


COMMENT DEFINICION SOURCE
REGION NUM=4 IX.LO=1 IX.HI=7 IY.LO=3 IY.HI=15 SILICON

COMMENT DEFINICION DRAIN
REGION NUM=5 IX.LO=28 IX.HI=35 IY.LO=3 IY.HI=15 SILICON


COMMENT ELECTRODO PUERTA SUPERIOR
ELECTRODE NUM=1 IX.LO=7 IX.HI=28 IY.LO=1 IY.HI=1

COMMENT ELECTRODO DE SOURCE
ELECTRODE NUM=2 IX.LO=1 IX.HI=1 IY.LO=3 IY.HI=15

COMMENT ELECTRODO DRAIN
ELECTRODE NUM=3 IX.LO=35 IX.HI=35 IY.LO=3 IY.HI=15

comment electrodo bulk
ELECTRODE NUM=4 IX.LO=1 IX.HI=35 IY.LO=45 IY.HI=45

DOP REGION=3 UNIFORM CONC=1E14 P.TYPE
DOP REGION=2 UNIFORM CONC=1E14 P.TYPE
DOP REGION=4 UNIFORM CONC=1e19 N.TYPE
DOP REGION=5 UNIFORM CONC=1e19 N.TYPE

CONTACT NU=1 N.pol
```

```
SYMBOLIC NEWTON CARRIERS=1
METHOD IT=100

comment LOG OUTF=CURRENTDOWN.DAT

SOLVE INIT

SOLVE V1=1  V4=0 V2=0 V3=1 OUT=SOL

METHOD 2ND TAUTO AUTONR
SOLVE V1=0
SOLVE V1=2 dt=1e-9 TSTOP=25E-9 RAMPTIME=10E-9
+ OUTF=UP1
```

Como ejemplo se muestran representaciones gráficas de la concentración de electrones tanto en representación bidimensional como tridimensional correspondientes al archivo de salida SOL.

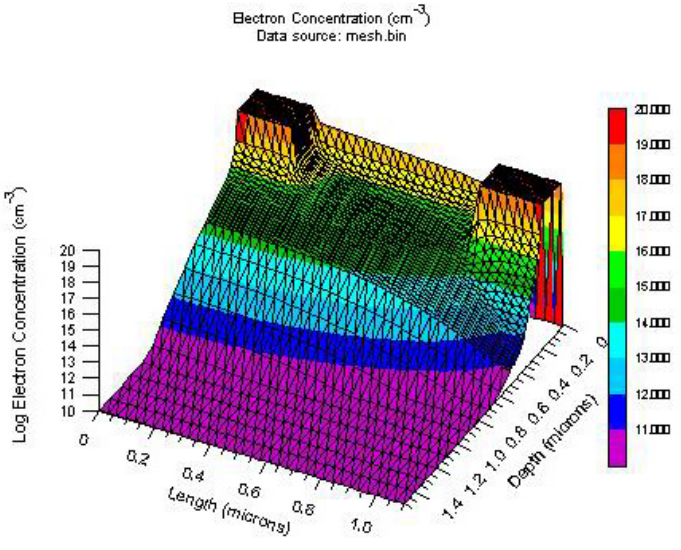


Figura 17.6.3 Representación 3D de la concentración de electrones en un MOSFET

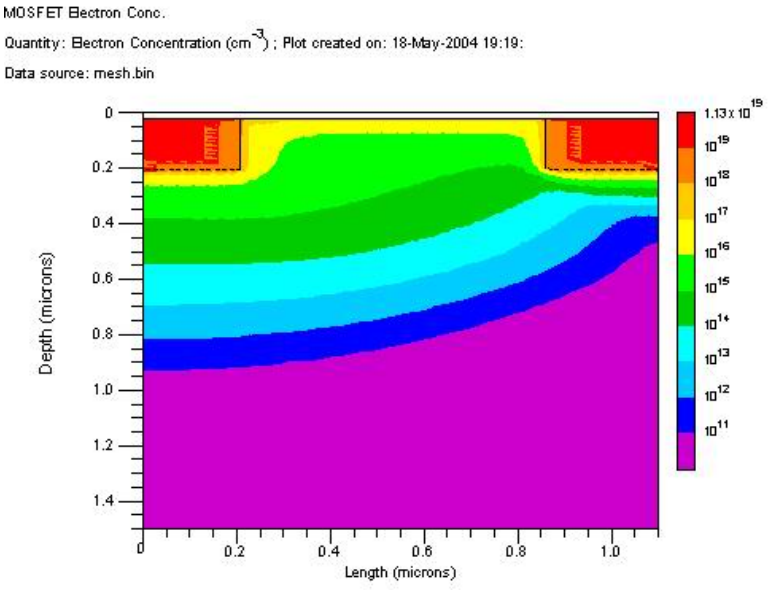


Figura 17.6.4 Diagrama de contorno para la representación 2D de la concentración de electrones en un MOSFET

17.7 Ejecución de PISCES

Los ejemplos mostrados en la sección anterior junto a una versión gratuita de PISCES, POSTMINI y los documentos de referencia de ambos programas pueden ser encontrados en el siguiente archivo.

Para realizar una simulación se debe abrir una ventana de DOS. Una vez descomprimido en un directorio el archivo adjunto, se ejecuta desde el directorio en que esté el archivo de PISCES la orden *pisces2 <nombre_de_archivo_a_simular>* la simulación se realizará de forma automática creándose los archivos de salida en el directorio actual si no se ha especificado lo contrario en el código.

RESUMEN

En este capítulo se ha realizado una descripción de la herramienta de simulación PISCES II y de la herramienta de presentación de datos POSTMINI. Una vez presentados los distintos modelos disponibles en PISCES se han explicado las distintas posibilidades de simulación y los comandos más importantes a incluir en la estructura del fichero de entrada. Tras enumerar las diferentes posibilidades de POSTMINI se han realizado como ejemplo dos simulaciones correspondientes a un diodo largo y a un transistor de efecto campo MOS (MOSFET).

REFERENCIAS

[1] Z. Yu, et al. *PISCES-2ET and its application subsystems*. Stanford University, 1994

[2] J. Faricelli. *POSTMINI User's Manual.* Alpha Semiconductor Technology, Hewlet Packard Company

[3] Stanford Technology CAD Home Page: http://www-tcad.stanford.edu/

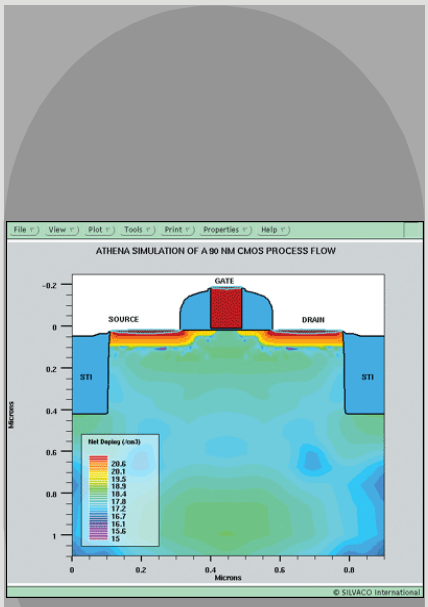
[4] Ports of Popular TCAD Software to Win32 http://home.comcast.net/~john.faricelli/tcad.htm

18

Capítulo

Simulación de Procesos con ATHENA

90 nm CMOS



Índice

Objetivos

- Presentar un simulador de procesos comercial
- Mostrar las diferentes capacidades de éste
- Describir su uso bajo el entorno DeckBuild
- Ilustrar la creación paso a paso de un transistor MOS

Palabras Clave

ATHENA.
SSUPREM.
ELITE.
OPTOLITH.
FLASH.
CMP.
CVP.
DeckBuild.
SSF.
TonyPlot.
Postmini.
Grid adaptativo.
Deposición de películas simples.
Difusión.
Oxidación.
Grabado.
Realajación del mallado.
Reflexión de estructuras.

18.1 Descripción de ATHENA

ATHENA es un simulador de propósito general que permite la simulación de procesos tecnológicos en semiconductores utilizando modelos bidimensionales basados en modelos físicos.

En la especificación del problema a resolver deben definirse los siguientes aspectos:

- Geometría inicial de la estructura que va a ser simulada
- La secuencia de procesos que se llevarán a cabo sobre el sustrato definido inicialmente
- Los modelos físicos que serán utilizados durante la simulación

ATHENA presenta una arquitectura modular integrada por las siguientes herramientas:

- SSUPREM4: Esta herramienta es utilizada para el diseño, análisis y optimización de estructuras semiconductoras basadas en Si simulando procesos típicos en esta industria tales como implantación iónica, difusión y oxidación.
- ELITE: Este módulo es un simulador de propósito general para topologías 2D que describe diferentes tipos de deposiciones y grabados utilizados en la fabricación de circuitos integrados.
- OPTOLITH: Permite la simulación de procesos de litografía óptica en general.
- FLASH: Módulo utilizado para el estudio de procesos en materiales compuestos tales como GaAs o SiGe.

18.2 Descripción de las posibilidades de ATHENA

Para cada uno de los distintos procesos tecnológicos a simular ATHENA ofrece distintas posibilidades en los modelos que se pueden utilizar. Las características más destacadas son las siguientes:

- Especificación de tiempos y temperatura de cocido.
- Modelos de pulido químico y mecánico (CMP) incluyendo pulidos duros, suaves, combinación de ambos y grabado isótropo.
- Modelos de deposiciones conformes, unidireccionales o bidireccionales. Modelos de metalización semiesférica, planetaria y cónica. CVD (Chemical Vapor Deposition). Efectos de difusión superficial y migración. Deposición balística y modelos definidos por el usuario.
- Difusión de impurezas en estructuras bidimensionales generales incluyendo difusión en todas las capas de materiales. Modelo completo de difusión de defectos

puntuales acoplados. Efectos de difusión retardada y oxidación mejorada. Modelos RTA (Rapid Thermal Annealing). Difusión de impurezas en polisilicio diferenciando entre el interior del grano y la frontera de éste.

- Simulación de epitaxia en 2D incluyendo efectos de autodopado.
- Capacidad de grabado geométrico. Grabado húmedo con perfil isótropo. Modelo RIE que combina componentes de grabado isótropo y direccional.
- Exposición sobre sustratos con topografía arbitraria, desenfoque y efectos de grandes aperturas numéricas.
- Formación de imágenes bidimensionales, bajo grandes aperturas numéricas y en medios aéreos. Inclusión de efectos de aberraciones en el sistema de hasta noveno orden. Capacidad de incluir máscaras de desplazamiento de fase y efectos de variación de transmitancia. Inclusión de efectos de fuentes de iluminación extensas.
- Implantaciones de tipo Pearson, doble Pearson y Gaussianas. Tablas de parámetros de implantación para bajas y altas energías mejorados. Simulación Monte Carlo de implantación para materiales amorfos y cristalinos.
- Modelos HCL y de presión mejorados para oxidación. Efectos de dependencia con la concentración de impurezas. Capacidad para simular la oxidación de estructuras con trincheras profundas y capas ONO. Modelos para oxidación y estiramiento simultáneos de regiones de polisilicio.
- Cálculos basados en modelos físicos incluyendo difusión simultánea de impurezas, segregación e inyección de defectos puntuales para procesos de silicidación.

18.3 Uso de ATHENA con DeckBuild

DeckBuild es un entorno gráfico interactivo que puede ser utilizado para generar los archivos de entrada para simulaciones de procesos o dispositivos, ejecutar simulaciones de modo interactivo o servir de interfase entre distintos simuladores.

Para iniciar ATHENA se debe llamar al siguiente comando UNIX `deckbuild -an` con lo que aparecerá una ventana similar a la de la Figura 18.3.1.

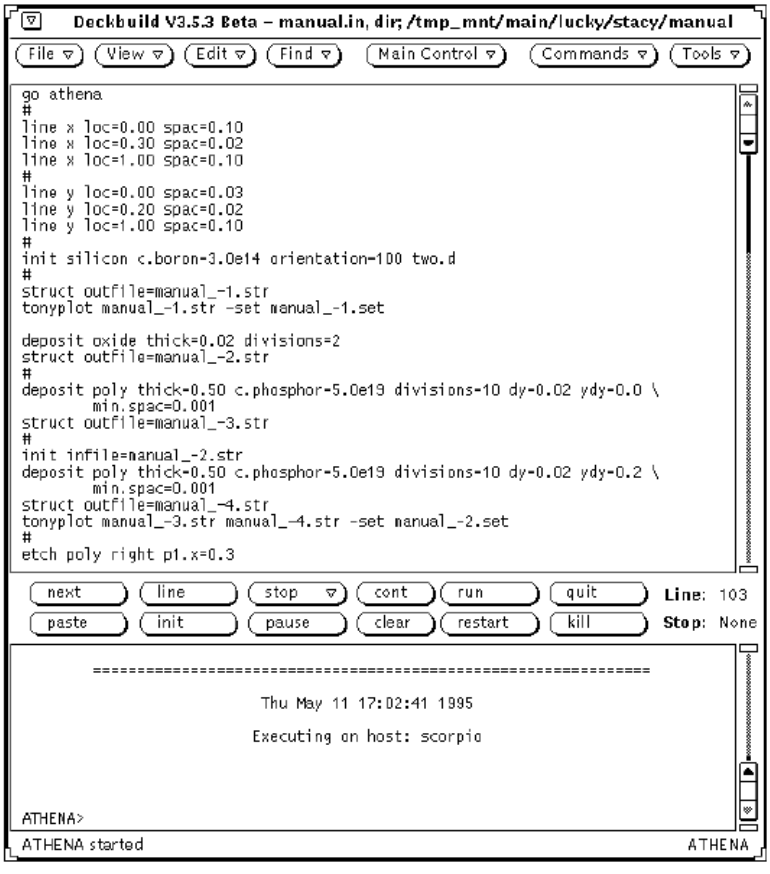


Figura 18.3.1 Ventana principal de DeckBuild.

En la ventana aparece información sobre la versión que se está utilizando y sobre los módulos instalados. Asimismo es posible encontrar un indicador de línea de comandos para introducir diferentes órdenes.

18.4 Creación de la estructura de un dispositivo

Las simulaciones realizadas con ATHENA se definen a partir de un fichero de entrada y dan lugar a un cierto número de ficheros de salida. A continuación se presentan los distintos pasos a realizar para obtener un fichero de entrada a partir del cual se pueda realizar una simulación.

- Desarrollar un buen mallado para la simulación
- Representar la deposición
- Representar el grabado geométrico
- Manipulación de las estructuras
- Grabar y cargar la información sobre las estructuras
- Comunicación con el simulador de dispositivos

Los ficheros de entrada pueden ser de dos tipos: archivos de configuración que contienen toda la información de los modelos y procesos que se pueden simular incluyendo los valores por defecto de los

parámetros que describen a éstos. El archivo de entrada creado por el usuario a partir de DeckBuild o de un editor ASCII estándar. Independientemente de la forma en que este haya sido creado en cada línea del fichero se encuentra una instrucción cuya estructura viene dada por el nombre de ésta y una serie de parámetros que modifican los efectos que dicha instrucción tiene sobre la simulación.

En cuanto a la información generada se tienen dos tipos principales de salida: En primer lugar la salida estándar provee la información resultante de la simulación a partir por ejemplo de instrucciones del tipo `PRINT.1D` y de las salidas creadas por ATHENA. En segundo lugar se tiene la salida estándar de errores que contiene todos los mensajes de error y advertencias que puedan ser generados durante el proceso de simulación.

Para el intercambio de información entre ATHENA y otros simuladores y herramientas se ha definido una estructura de archivo denominada SSF (Standard Structure File) cuyo formato es universal para diferentes herramientas de Silvaco. La instrucción `STRUCTURE` crea un SSF que contiene el mallado, información sobre la solución, los modelos y otras informaciones relevantes.

El SSF resultante puede ser utilizado para reiniciar la estructura y continuar la simulación de procesos desde el punto en que se quedaron, para simular el dispositivo resultante en un simulador eléctrico como puede ser ATLAS o PISCES o para representar los resultados con una herramienta gráfica como TonyPlot o Postmini.

A continuación se presentan los distintos pasos que se deben seguir para crear una estructura a partir del editor DeckBuild. En este proceso no se va a crear ninguna estructura real, sino que se va a describir el uso de distintas instrucciones para explorar las distintas posibilidades ofrecidas que luego deben ser usadas para la creación de un dispositivo real.

Definición de un mallado rectangular

Para la definición del mallado se debe marcar en el menú Commands y, una vez desplegado éste, se elige la opción Mesh Define. En este momento se puede especificar el mallado inicial. La correcta especificación de éste es fundamental para obtener una simulación lo más precisa posible y está directamente relacionada con el número de nodos. Es recomendable definir un mallado fino en las zonas en que se vaya a producir una implantación iónica, en las uniones p-n o donde la iluminación vaya a cambiar las propiedades de la resina fotosensible. El número máximo de nodos que se pueden utilizar es de 20.000 aunque lo normal es utilizar muchos menos.

Para crear un mallado uniforme de 1 micra por 1 micra primero se debe seleccionar el campo Location introduciendo el valor 0.0 y a continuación seleccionar el campo Spacing dándole el valor 0.1. Haciendo click sobre el botón Insert. Del mismo modo se debe repetir la operación

para una línea situada en 1.0 con el mismo espaciado anterior. Una vez hecho esto, se selecciona la dirección Y y se repite la operación. Entonces debería aparecer el Mesh Define Menu como se muestra en la siguiente figura,

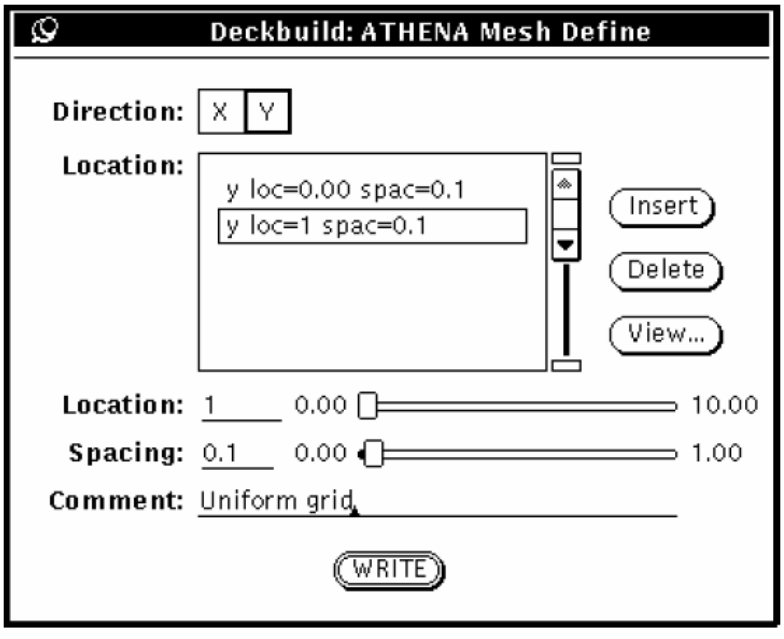


Figura 18.4.1 Menú Mesh Define

En este momento es posible escribir la información en el fichero de entrada, aunque es recomendable visualizar el mallado generado utilizando la opción **View**.

Un mallado uniforme como el que se acaba de generar resulta ser ineficiente para realizar simulaciones complejas. En primer lugar se va a mejorar el mallado en la dirección Y. Si la opción de grid adaptativo está deshabilitada se debe realizar un estudio a priori de los procesos que se van a llevar a cabo.

Por ejemplo, si se va a realizar una implantación de B de 60 keV se sabe que el máximo del perfil implantado estará aproximadamente en 0.2 micras de profundidad. Para mejorar el mallado en esta zona basta con añadir una nueva línea en la dirección Y en la posición 0.2 y con un espaciado de 0.02. De esta forma se genera un mallado cuya separación va disminuyendo gradualmente de 0.1 a 0.02. Debido a que el espaciado original era de 0.1, sólo existen 3 líneas entre la superficie y el punto de máxima concentración tras la implantación. Si se hace click sobre la línea situada en Y = 0.0 es posible cambiar el espaciado a 0.03 por ejemplo, de forma que quede más fino en esa zona.

Para mejorar el mallado en la dirección X se deben tener en cuenta dos factores. En primer lugar uno debe asegurarse de que se obtiene una buena resolución bidimensional bajo los límites de las futuras

máscaras y en segundo lugar que las líneas verticales están situadas a lo largo de los bordes de las máscaras.

Si se quiere construir la mitad de una estructura MOS de 0.6 micras con el centro de la puerta situada en X=0 se debe añadir una línea adicional en X=0.3 y el espaciado debe resultar lo suficientemente bueno como para obtener una buena resolución lateral en los implantes de fuente y drenador.

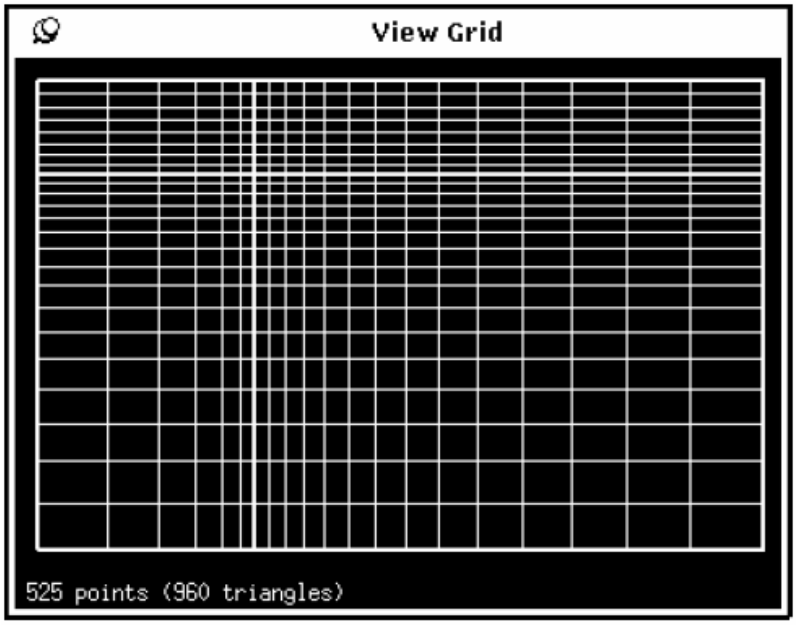


Figura 18.4.2 Mallado gradual

Cuando el mallado tenga la disposición que se esperaba, se puede escribir la definición de este en el archivo de entrada pulsando sobre **Write**. El archivo queda de la siguiente forma:

```
GO ATHENA
# NON-UNIFORM GRID
LINE X LOC=0.00 SPAC=0.1
LINE X LOC=0.3 SPAC=0.02
LINE X LOC=1 SPAC=0.1
LINE Y LOC=0.00 SPAC=0.03
LINE Y LOC=0.2 SPAC=0.02
LINE Y LOC=1 SPAC=0.1
```

La primera línea se denomina instrucción de *autointerface* indicando que la siguiente parte de código debe ser ejecutada utilizando ATHENA.

Definición del Sustrato Inicial

Hasta ahora sólo se ha definido la estructura que servirá de base para la simulación. El siguiente paso consiste en definir el sustrato sobre

el que se realizarán los procesos posteriores. Para ello se debe seleccionar la opción **Mesh Initialize** con lo que aparecerá el siguiente menú:

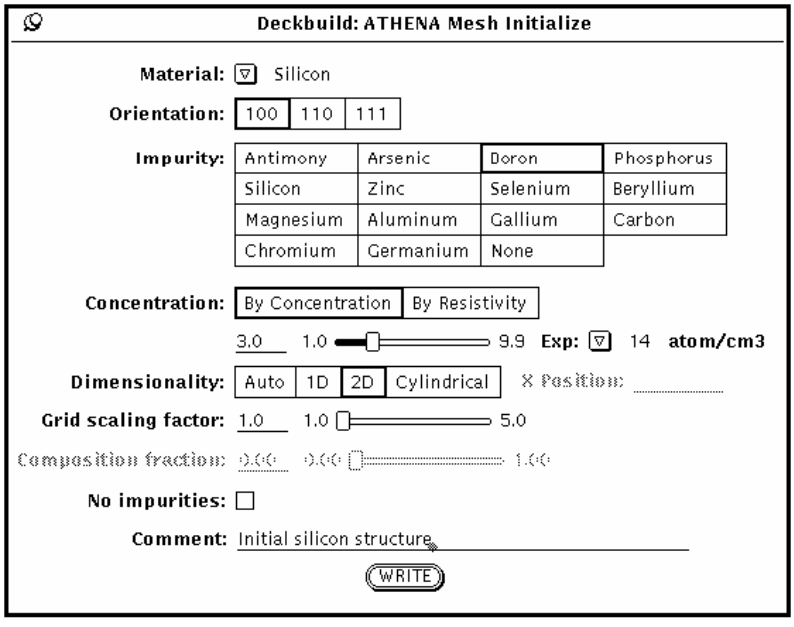


Figura 18.4.3 Mesh Initialize

En este menú es posible definir diferentes parámetros del sustrato como pueden ser el tipo de material, la orientación cristalina o el tipo y concentración de impurezas que se van a implatar. De nuevo usando el botón **Write** la información es transferida al archivo de inicio:

```
# INITIAL SILICON STRUCTURE
INIT SILICON C.BORON=3.0E14 ORIENTATION=100 TWO.D
```

Ahora se debe ejecutar ATHENA pulsando **Run** para obtener la estructura inicial.

DeckBuild genera un archivo llamado histoy01.str que permite realizar simulaciones del tipo “what if” o visualizar en cualquier momento la estructura. Una vez realizado este proceso, se obtiene un sustrato listo para sufrir diferentes tipos de procesos.

Deposición de Películas simples

La deposición conforme puede ser utilizada para generar estructuras multicapa. Este modelo es el más simple y puede ser utilizado cuando la forma exacta de la lámina depositada no es crítica. Para realizar un paso de deposición se debe seleccionar en el menú **Process>Deposit>Deposit...**

Este tipo de deposición está seleccionada por defecto, a continuación se seleccionará una deposición de óxido de un espesor de

0.02 micras. Es recomendable que haya al menos dos líneas de mallado en la zona crecida aunque si se quiere representar mejor la forma final de la capa depositada se deben usar más. El mallado está controlado por el parámetro **Grid Specification** en el menú de deposición. Una vez seleccionado el número de capas a incluir se pulsa de nuevo **Write** y la ventana principal se actualizará como sigue,

```
# GATE OXIDE DEPOSITION
DEPOSIT OXIDE THICK=0.02 DIVISIONS=2
```

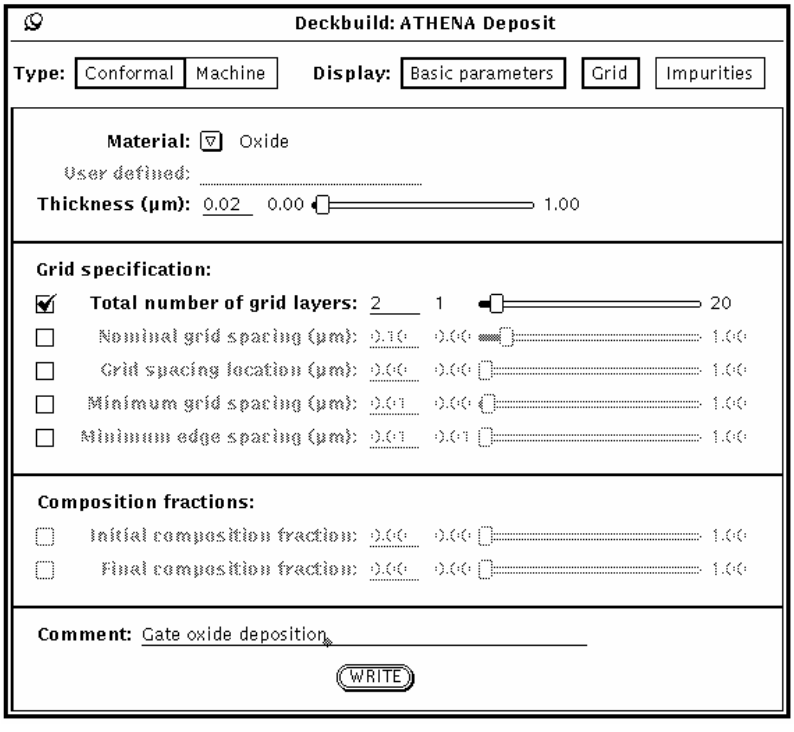


Figura 18.4.4 Menú de configuración para una deposición de óxido

El siguiente paso será la deposición de una capa de polisilicio dopado con fósforo de 0.5 micras de espesor. Se debe elegir dicho material y la concentración puede fijarse en el cuadro de diálogo. Es posible seleccionar un mallado más fino para esta zona fijando el número de capas en 10 con un espaciado nominal de 0.02 micras y localizándolo en la posición 0.0 correspondiente a la superficie. El código queda de la siguiente forma:

```
DEPOSIT POLY THICK=0.5 C.PHOSPHOR=5.0E19 DIVISIONS=10 \
DY=0.02 YDY=0.0 MIN.SPACING=0.001
```

Pulsando el botón **Cont** se continúa la simulación de ATHENA de forma que se crea la estructura de tres capas definida con anterioridad. El resultado puede observarse en la Figura 18.4.5

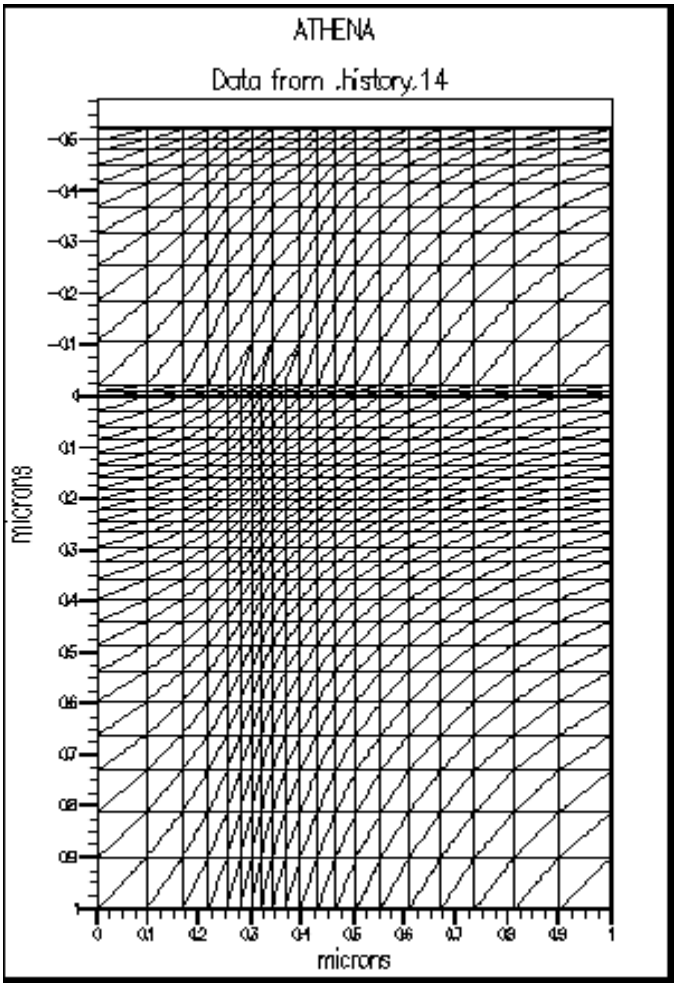


Figura 18.4.5 Estructura generada tras la deposición de óxido y polisilicio

Grabado Geométrico

El siguiente paso en el proceso consiste en definir la puerta de polisilicio. Para ello se debe seleccionar el proceso de grabado de la forma **Process - Etch - Etch...** En el ejemplo siguiente se tiene una puerta de polisilicio con el borde en X=0.3 micras con el centro situado en X=0. Para ello se debe seleccionar **Right** en el apartado **Geometrical type** y dar el valor 0.3 a **Etch location (**Figura 18.4.6**)**.

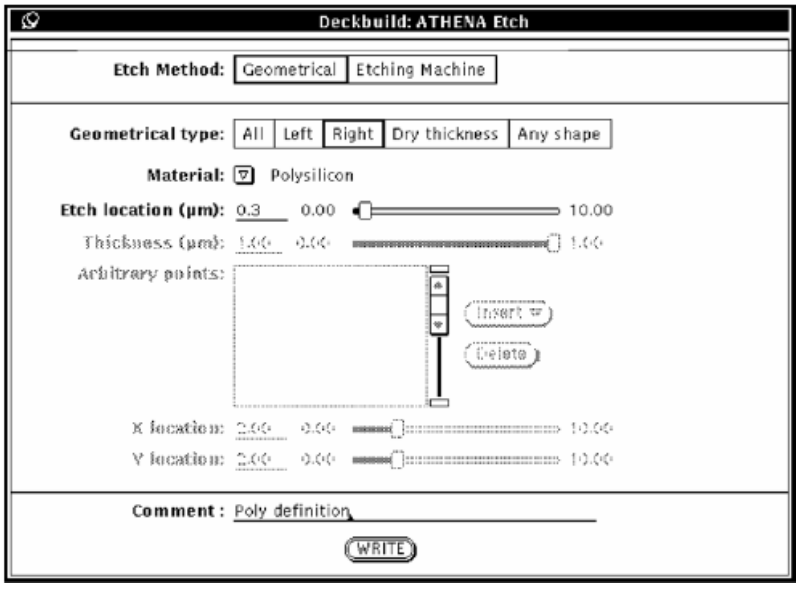


Figura 18.4.6 Menu de Grabado

Una vez hecho ésto se obtiene el siguiente código

```
# POLY DEFINITION
ETCH POLY RIGHT P1.X=0.3
```

La estructura creada puede ser observada en la parte izquierda de la Figura 18.4.8. También es posible definir una forma arbitraria para el grabado geométrico utilizando la opción **Any shape**, como puede ser observado en la Figura 18.4.7

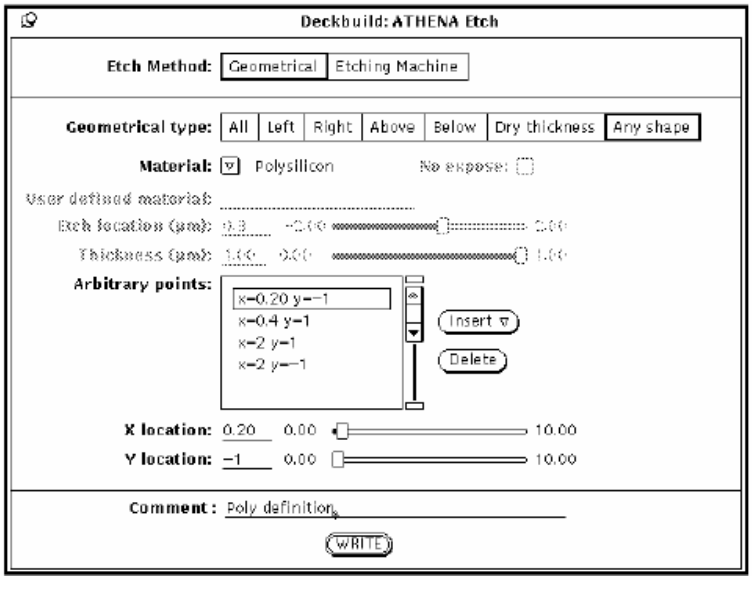


Figura 18.4.7 Menú de Grabado a medida

El código generado para este caso es el siguiente:

```
# POLY DEFINITION
ETCH POLY START X=0.2 Y=-1
ETCH CONT X=0.4 Y=1
ETCH CONT X=2 Y=1
ETCH DONE X=2 Y=-1
```

A continuación se muestra el resultado del grabado para las opciones geométrica normal (izda) o a medida (dcha)

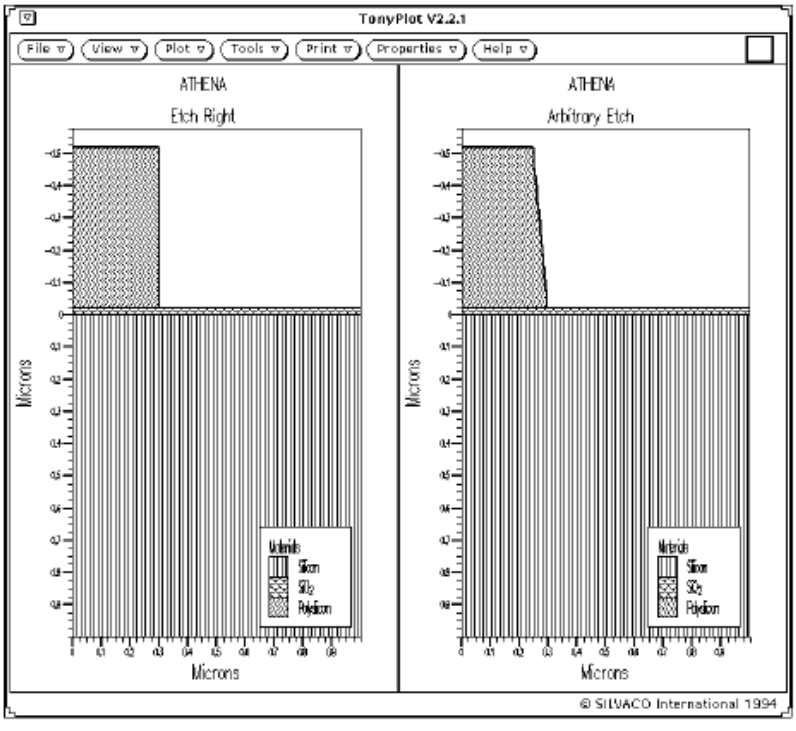


Figura 18.4.8 Resultado tras realizar grabado geométrico normal (izda) o definido por el usuario (dcha)

Reducción del número de puntos del mallado en áreas no esenciales

Como se ha comentado con anterioridad, la calidad del mallado es de extrema importancia para la consecución de una buena simulación. El mallado generado hasta ahora ha permanecido intacto en las zonas que no han sufrido procesos que afecten al grid. Existe una capacidad denominada relajación del mallado (Grid Relax) que permite incrementar el espaciado de los puntos en ciertas zonas del dispositivo. Esta opción resulta muy útil desde dos puntos de vista principales.

En primer lugar las zonas de mallado muy fino se propagan por toda la estructura cuando sólo resultan necesarios en ciertas zonas. En segundo lugar reduciendo el número de puntos se consigue reducir de una

forma apreciable el tiempo de simulación sin afectar a la precisión de la simulación en sí. Estos parámetros pueden ser seleccionados en el correspondiente menú mostrado en la Figura **18.4.9**

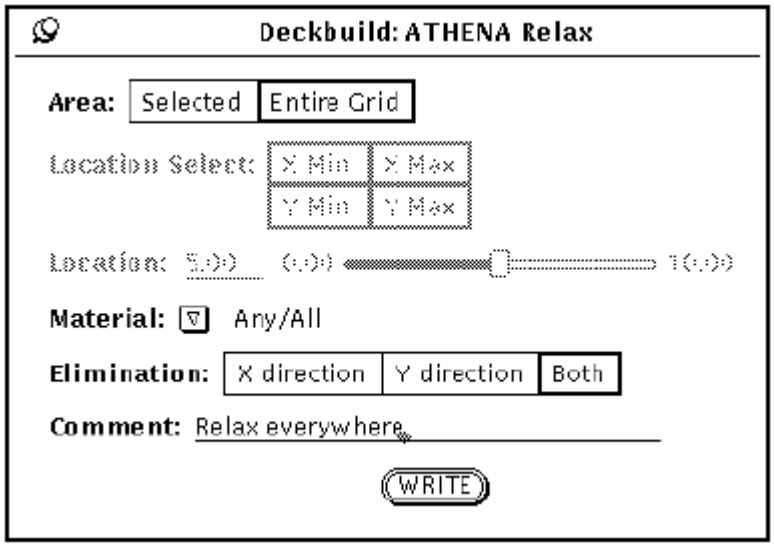


Figura 18.4.9 Menú de parámetros de relajación del mallado

A continuación se muestran los efectos de la aplicación de esta característica a los procesos anteriores y en distintas zonas de la estructura. El archivo de entrada será modificado de la siguiente forma según la zona de la estructura en la que se cambie el mallado:

```
# RELAX EVERYWHERE
RELAX DIR.X=T DIR.Y=T

# RELAX LOWER HALF OF THE STRUCTURE
RELAX  X.MIN=0.00  X.MAX=1.00  Y.MIN=0.3  Y.MAX=1.00  DIR.X=T
DIR.Y=T

#  RELAX  LOWER  HALF  ONLY  IN  X-DIRECTION  RELAX  X.MIN=0.00
X.MAX=1.00 Y.MIN=0.3
Y.MAX=1.00 DIR.X=T DIR.Y=F
```

Quedando la estructura como se muestra en la Figura 18.4.10

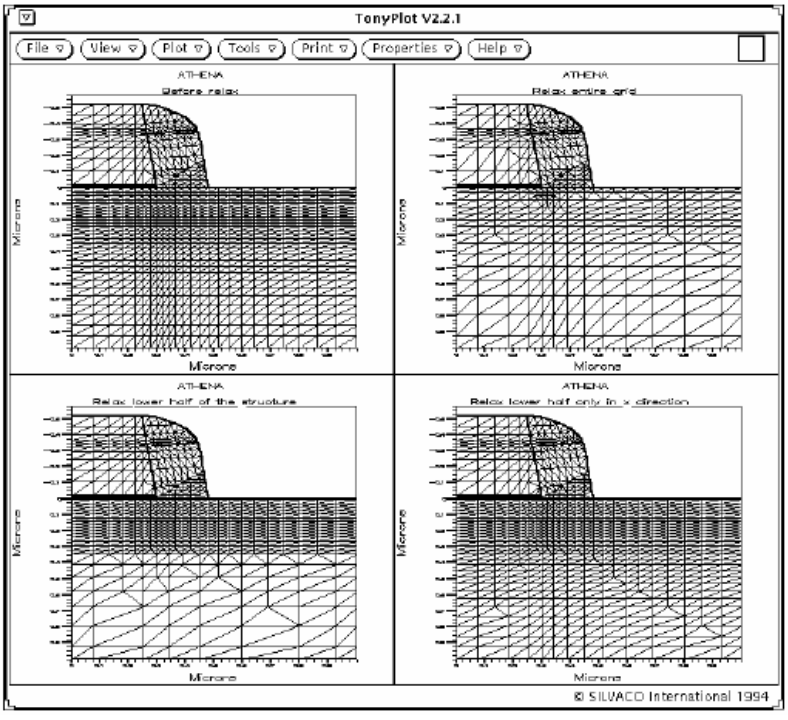


Figura 18.4.10 Estructura resultante tras realizar relajaciones del mallado con distintos parámetros

Básicamente el algoritmo está diseñado para seguir las siguientes pautas de comportamiento:

- Sólo puede ser utilizado con mallado rectangular
- Nunca se eliminan puntos adyacentes a la frontera de una región
- El área a relajar debe tener, al menos, 5 x 5 puntos
- No se permite formar triángulos obtusos

Reflexión de Estructuras

Hasta este punto se ha realizado el equivalente a la mitad de una estructura MOSFET. En algún momento antes de definir los contactos o exportar la estructura a un simulador de dispositivos se debe obtener la estructura completa. En general esto se debe de realizar cuando la estructura vaya a dejar de ser simétrica.

Para realizar esto con la estructura que se ha ido creando hasta ahora se debe seleccionar **Mirror** en el menú de instrucciones y se añadirá lo siguiente a la línea de comandos:

```
STRUCT MIRROR LEFT
```

El resultado de esta operación es el mostrado en la siguiente figura:

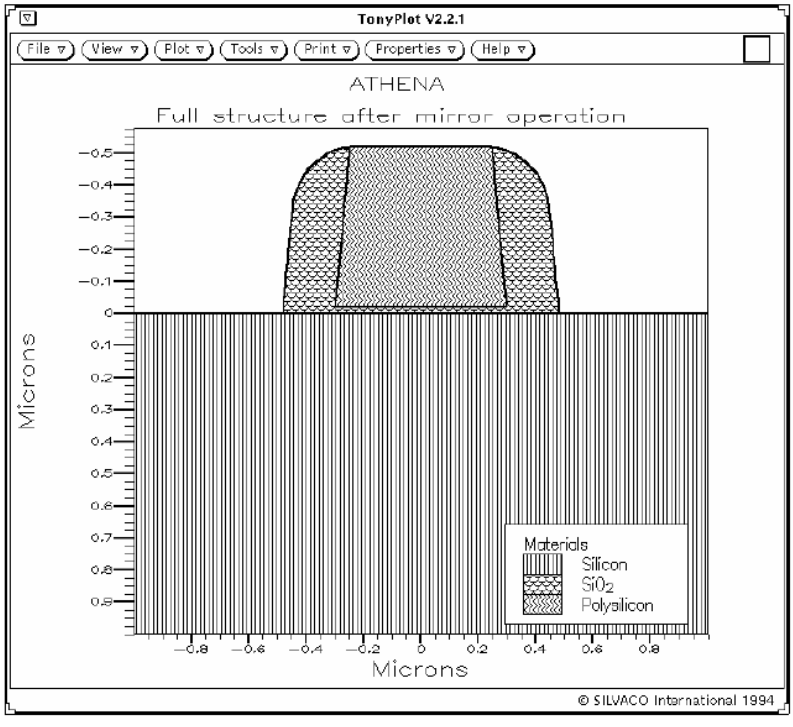


Figura 18.4.11 Estructura final MOSFET con los electrodos definidos

Especificación de Electrodos

Normalmente, la estructura que se quiere crear va a ser utilizada en un simulador de dispositivos para ser caracterizada eléctricamente. Para ello es necesario definir los electrodos que servirán de contactos para aplicar las distintas polarizaciones externas.

ATHENA puede asignar un electrodo a cualquier estructura de siliciuro, metal o polisilicio. Por ejemplo si se deposita una lámina de aluminio de 0.1 micras de espesor en toda la estructura y después se realiza un grabado en parte de ella utilizando la opción **Any Shape**

```
DEPOSIT ALUMIN THICK=0.1
ETCH ALUMINUM START X=-0.8 Y=-20
ETCH CONT X=-0.8 Y=20
ETCH CONT X=0.8 Y=20
ETCH DONE X=0.8 Y=-20
```

El siguiente paso consiste en definir los electrodos en el menú **Electrode** en donde se debe especificar la posición y el nombre de éste de forma que se agregarán las siguientes tres líneas

```
ELECTRODE NAME=SOURCE X=-0.9
```

```
ELECTRODE NAME=DRAIN X=0.9
ELECTRODE NAME=GATE X=0.0
```

Para especificar el electrodo del sustrato se debe elegir en el mismo menú un electrodo de tipo Backside de forma que se añadirá la siguiente instrucción

```
ELECTRODE NAME=BACK BACKSIDE
```

Tras aplicar estos pasos se obtiene el resultado mostrado en la Figura 18.4.12.

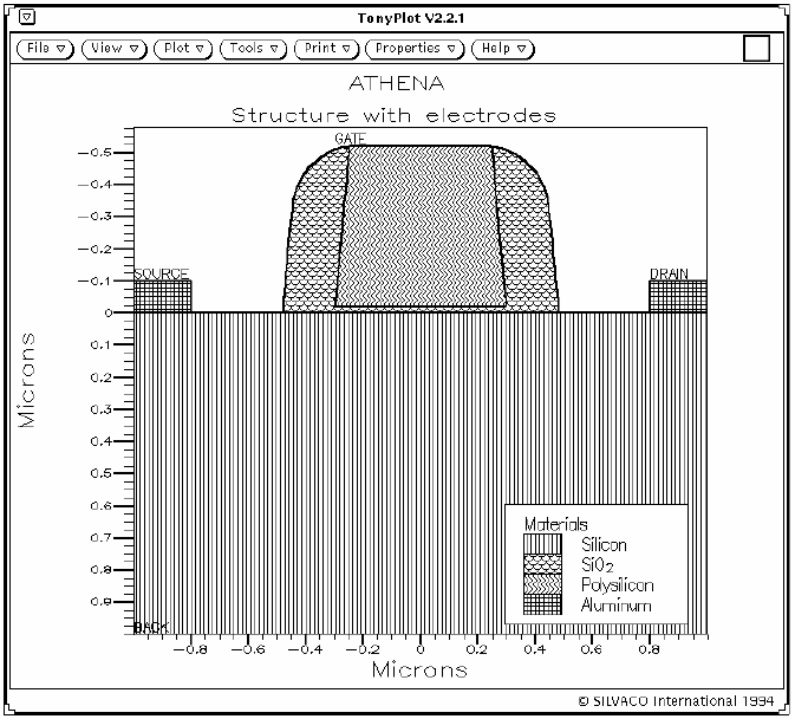


Figura 18.4.12 Estructura final MOSFET con los electrodos definidos

De esta forma se obtiene la estructura final que se quería simular. A partir de ella, y una vez grabada, es posible generar un archivo de salida con la caracterización del dispositivo para que éste sea caracterizado eléctricamente en otro simulador como puede ser PISCES o ATLAS.

Finalmente, a modo de ejemplo se muestran varios ejemplos de dispositivos creados con ATHENA que pueden ser encontrados en el sitio web de Silvaco www.silvaco.com

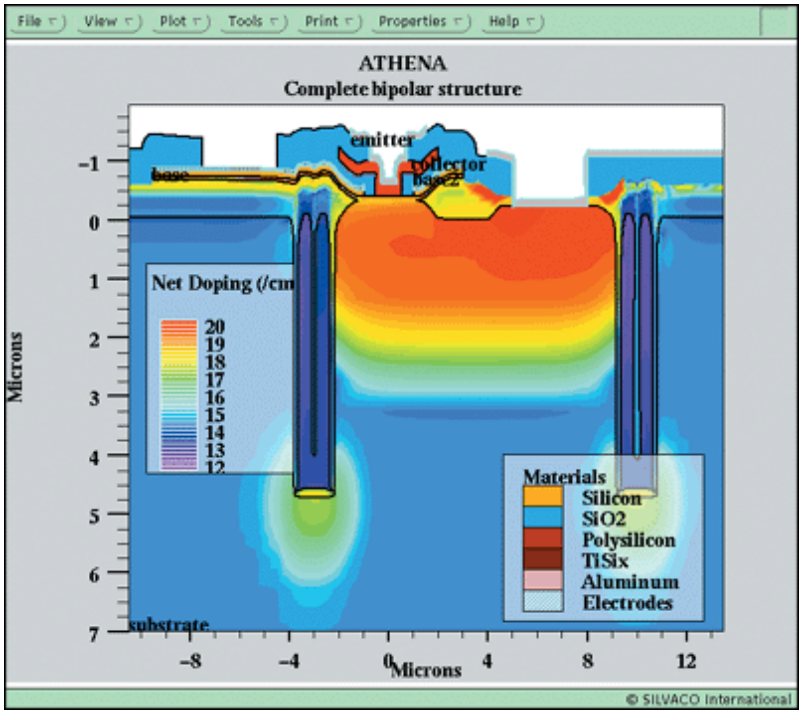


Figura 18.4.13 Estructura de Transistor Bipolar

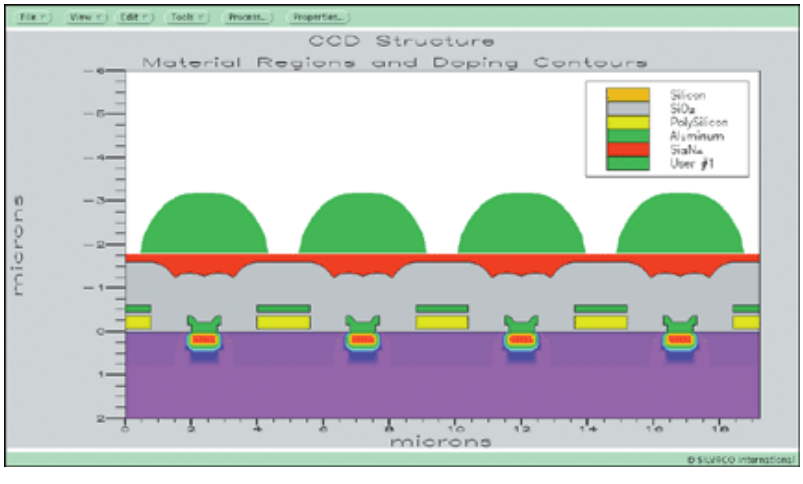


Figura 18.4.14 Estructura de dispositivos de carga acoplados (CCD)

RESUMEN

En este capítulo se ha presentado el simulador de procesos ATHENA de la empresa Silvaco utilizado ampliamente tanto en la industria como para propósitos de investigación. Tras una descripción de las capacidades de la herramienta de simulación, se han comentado los distintos modelos que incluye y la integración en el entorno visual DeckBuild para UNIX. Finalmente se han descrito los procedimientos para crear una estructura dada partiendo de la definición del mallado hasta la creación de los electrodos utilizando como guía el proceso paso a paso de definición de una estructura MOSFET.

REFERENCIAS

[1] *ATHENA User's Manual. Silvaco Internacional*. Santa Clara, CA. 2002

[2] Silvaco Internacional, http://www.silvaco.com